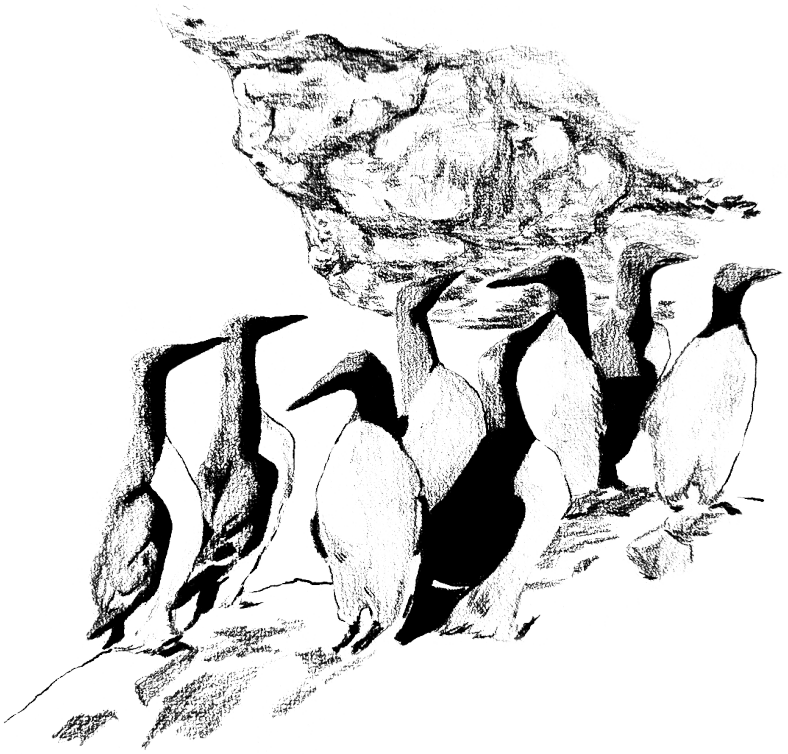

Multi-modal Machine Listening for Monitoring Guillemot Behaviour at Stora Karlsö

Automated acoustic detection of prey delivery using pre-trained bioacoustic models

Degree project report in Complex Adaptive Systems

Léa Le Gallo & Guðmundur Andri Guttormsson

DEPARTMENT OF PHYSICS, UNIVERSITY OF GOTHENBURG
DEPARTMENT OF PHYSICS AND ASTRONOMY, CHALMERS

Gothenburg, Sweden 2026
www.gu.se www.chalmers.se

Degree project report 2026

Multi-modal Machine Listening for Monitoring Guillemot Behaviour at Stora Karlsö

Automated acoustic detection of prey delivery using pre-trained bioacoustic models

Léa Le Gallo & Guðmundur Andri Guttormsson

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics - University of Gothenburg
Department of Physics and Astronomy - Chalmers University of Technology
Gothenburg, Sweden 2026

Multi-modal Machine Listening for Monitoring Guillemot Behaviour at Stora Karlsö
Automated acoustic detection of prey delivery using pre-trained bioacoustic models
Léa Le Gallo & Guðmundur Andri Guttormsson

Supervisor: Olof Mogren, Senior Researcher at RISE
Examiner: Mats Granath, Professor at the Department of Physics, University of Gothenburg

Degree project report 2026
Department of Physics and Astronomy, Chalmers
Department of Physics, University of Gothenburg
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000 & +46 31 786 00 00

Cover: Graphite on paper — digitised and digitally processed in Photoshop. Art piece by Cloé Le Gallo, inspired by a photograph of guillemots taken at Stora Karlsö by Guðmundur Andri Guttormsson.

Cloé's LinkedIn

Typeset in LaTeX
Gothenburg, Sweden 2026

Multi-modal Machine Listening for Monitoring Guillemot Behaviour at Stora Karlsö
Automated acoustic detection of prey delivery using pre-trained bioacoustic models
Léa Le Gallo & Guðmundur Andri Guttormsson
Department of Physics
Chalmers University of Technology & University of Gothenburg

Abstract

Automated analysis of seabird behaviour is challenging due to the complexity of dense colony environments and the practical limits of manual annotation over continuous recordings. This thesis presents a machine listening pipeline trained on acoustic recordings collected at the Karlsö AukLab on Stora Karlsö, Sweden, during the 2025 breeding season, distinguishing three classes: arrivals with fish, arrivals without fish, and non-events (background colony noise). Accurate classification of these events is directly linked to chick provisioning rates and breeding success, making it a valuable tool for long-term ecological monitoring of seabird populations. A labelled dataset was constructed from YOLO detections on colony camera footage and verified through manual inspection, replacing the labour-intensive manual annotation previously performed by researchers at the lab. Audio segments were extracted from DPA microphones and preprocessed. Three frozen pre-trained CNN models (BirdNET, PANN, and Perch) were evaluated as feature extractors with lightweight classifier heads, trained on the manually curated dataset and evaluated on both the curated and an automatically extracted dataset. BirdNET achieved the strongest performance (F1-score 82%), with fish arrivals being the most acoustically separable class. Embedding inspection revealed ecologically interpretable acoustic signatures, with chick vocalisations and adult call patterns playing a key discriminative role. Incorporating weather data further improved classification performance. The model generalised well to a held-out, non-curated dataset, demonstrating robustness beyond the manually verified training distribution. These results demonstrate that event-driven multimodal annotation combined with pre-trained bioacoustic embeddings enables reliable, scalable acoustic monitoring of prey delivery in dense seabird colonies, offering a practical path toward automated and passive long-term ecological monitoring.

Keywords: bioacoustics, embeddings, pre-trained models, guillemot, machine listening, seabird monitoring, prey delivery, BirdNET, PANN, Perch, multimodal, event detection.

Acknowledgements

First and foremost, we want to thank our supervisor, Olof Mogren, for his guidance throughout the thesis and for always being approachable and genuinely nice to work with. A big thank you also to the rest of the RISE team and the master's students we got to work with, John, Delia, Ida, Kamal, and Vineeth, for all the help and good discussions. Being part of a broader group of parallel thesis projects was genuinely motivating, and sharing ideas and progress along the way made the work better.
We are also incredibly grateful to Jonas Hentati-Sundberg and the team at Stora Karlsö, Aron, PA, Saga and Sara, for welcoming us to the island and allowing us to be part of their world for a couple of days. A special thanks to Ida as well, for driving us there and letting us stay at her place after the trip. Coming to the island and helping with setting up the cameras and microphones was one of the highlights of the thesis: great nature, great food, great music and thousands of guillemots.
This was genuinely a one-of-a-kind experience and we feel really lucky to have worked on a thesis that took us to a remote Swedish island, introduced us to a colony of seabirds, and resulted in some genuinely exciting machine listening research.

Léa Le Gallo & Guðmundur Andri Guttormsson, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
DPA	Danish Pro Audio (microphone brand)
F1	F1-score (harmonic mean of precision and recall)
GPS	Global Positioning System
MLP	Multi-Layer Perceptron
MFCC	Mel-Frequency Cepstral Coefficient
NTP	Network Time Protocol
PANN	Pretrained Audio Neural Network
PCEN	Per-Channel Energy Normalisation
PCA	Principal Component Analysis
RMS	Root Mean Square
SGD	Stochastic Gradient Descent
SL	Supervised Learning
SNR	Signal-to-Noise Ratio
SSL	Self-Supervised Learning
SVM	Support Vector Machine
TNN	Temporal Neural Network
UMAP	Uniform Manifold Approximation and Projection
ViT	Vision Transformer
YOLO	You Only Look Once

Contents

List of Figures

List of Tables

1
Introduction

1.1 Study site and monitoring infrastructure

This master's thesis is part of the AukLab Audio project at Stora Karlsö, a long-term research site for studying common guillemots (*Uria aalge*) in the Baltic Sea. The work contributes to the Baltic Seabird Programme and is based on a large multi-channel, multi-modal dataset comprising synchronised audio recordings, video footage, weather measurements, and other sensor data, collected during the breeding season from May to August.

The Karlsö AukLab was established in 2008 as an artificial nesting site designed to facilitate close monitoring of guillemots and the deployment of technological equipment.

AukLab juts from Stora Karlsö's cliff edge 30 meters above the Baltic, it consists of artificial ledges made out of rocks to resemble the surrounding natural environment. The 2025 set up of the cameras and microphones is further explained in section 2.5.2.

Since then, the colony has been continuously studied using cameras and sensors, providing insights into how environmental factors such as storms, heat waves, and predator presence (e.g., eagle fly-overs) influence breeding success [1].

More recently, the monitoring system has been extended with high-resolution audio, including a continuous eight-channel recording system and microphones integrated into IP cameras. Given that guillemots are highly vocal, producing sounds associated with courtship, territorial interactions, and feeding, this addition enables simultaneous acoustic and visual observation of colony behaviour.

The Baltic Sea is a sensitive marine ecosystem affected by climate change and eutrophication [2]. Stora Karlsö hosts one of the largest guillemot colonies in the Baltic Sea and is therefore a key site for long-term ecological monitoring. The development of automated, non-invasive methods for analysing this type of multimodal data is central to the project, enabling large-scale and continuous observation of seabird populations.

1.2 Biological motivation

A central focus in seabird research is understanding the drivers of breeding success, particularly the relationship between prey availability and reproductive performance. Seabirds such as the common guillemot serve as ecological indicators, as

their breeding success and behaviour reflect changes in prey availability and environmental conditions.
Prey delivery rates are a key component of this relationship. Low fish delivery can lead to reduced breeding success and may indicate broader ecological changes, such as declining fish stocks or environmental factors affecting foraging efficiency, for example through changes in prey visibility or flight conditions.
This thesis focuses on prey delivery as a measurable indicator of foraging success and ecosystem conditions. In particular, we focus on arrival events on the ledge, distinguishing between birds arriving with fish and without fish. These events provide a structured framework for analysing behaviour, as they allow us to examine how guillemots respond to feeding-related contexts. By linking these events to acoustic and visual signals, it becomes possible to infer behavioural states such as feeding, begging, and social interactions.

1.3 Automation and multimodal analysis

Recent advances in computer vision and machine learning have enabled new ways of extracting information from large-scale ecological data, particularly in systems such as the Karlsö AukLab. These approaches allow the automatic derivation of biologically relevant features directly from raw video, supporting tasks such as estimating chick growth, tracking individual birds on the ledge, reading identification rings using object detection and optical character recognition, reconstructing 3D pose, detecting prey delivery events, and quantifying environmental stress such as heat exposure.
A key advantage of these methods is their ability to generate datasets at a temporal and spatial resolution that would be impossible to obtain manually. This opens up new possibilities for analysing behaviour and identifying patterns in large, complex ecological systems. However, the scale of such data also makes manual processing infeasible. Given that the camera system at Stora Karlsö produced approximately 7.8 billion image frames in 2023, a human analyst would require around 247 years to review the data at a rate of one frame per second. This highlights both the scale of the data and the necessity of automated methods for efficient analysis [1].
Bioacoustics is valuable in ecology and conservation because it allows for non-invasive monitoring, and several important biological applications have been identified for automatic detection. Such methods can be used to estimate species occupancy and density, revealing where species occur and how common they are, which is essential for conservation planning. Spatial analyses benefit from mapping animal activity across landscapes to understand habitat use and movement; for example, bioacoustic monitoring can track migratory species over large ranges in a far less invasive way than individual tagging and without relying on individual survival. Automatic detection also supports the study of species characteristics, such as call types, which are central to behavioural ecology. In addition, vocal differences across regions or social groups provide insight into population structure and cultural or historical patterns, while individual vocal traits enable non-invasive identification, population monitoring, and welfare assessment [3].
While computer vision has enabled detailed analysis of visual behaviour, bioacous-

tics provides a complementary, non-invasive perspective on animal activity. In this context, the project explores machine listening methods for analysing complex acoustic scenes. The core challenge is to replace manual annotation with computational methods while maintaining accuracy, ideally using approaches that require limited labelled data. By combining audio, video, and environmental data, the aim is to develop models capable of detecting and classifying events such as arrivals with and without fish under real-world colony conditions.

1.4 Research gap and aim of the thesis

Despite recent progress in bioacoustics and computer vision for ecology, most methods still treat audio and video separately. This makes it difficult to capture complex, multimodal behaviours such as prey delivery in dense seabird colonies. Current approaches also rarely model behaviour as structured events over time or link signals to clear biological contexts, which is especially challenging in noisy, high-density environments. In addition, modern ecological datasets are too large for manual analysis, while many machine learning methods still depend on large labelled datasets, limiting practical use.

This thesis addresses these limitations by leveraging the large multimodal dataset of the AukLab guillemot colony and focusing on event-based modelling of arrival behaviour. In particular, it distinguishes between arrivals with and without prey by first extracting events from video analysis, using these events to construct a three-class audio dataset, and then classifying the corresponding audio segments using frozen CNN models (BirdNET, PANN, and PERCH). Weather data is also included and combined with the learned embeddings in the classifier to evaluate whether it improves classification performance.

Overall, the goal is to develop computational methods for analysing multimodal ecological data that improve the detection and interpretation of behaviour in complex colony environments. This contributes to a better understanding of guillemot behaviour and supports scalable, non-invasive monitoring of seabird populations through machine listening approaches.

1.5 Research questions

RQ1: Can guillemot arrival events and whether birds arrive with fish or without fish be automatically detected from acoustic recordings, using video and sensor data for event annotation and dataset construction?

RQ2: How do guillemot vocalisations differ between arrival events with fish and without fish?

RQ3: Does incorporating multimodal information (video, audio, and environmental data) improve the detection and classification of arrival-related events compared to audio-only models?

RQ4: Can weakly supervised or self-supervised learning reduce the need for manual annotation while maintaining reliable event detection and classification?

RQ5: How well do different machine learning models perform in detecting and classifying arrival events under noisy, real-world colony conditions?

2
Background

2.1 Bird bioacoustics

Birds are the most commonly studied taxon in bioacoustics, due to their diverse and easily detectable vocalisations. Their calls provide valuable information for monitoring populations, understanding behaviour, and assessing ecosystem health. Bioacoustic data on birds are widely available, making them an ideal focus group for developing and evaluating automated detection and classification methods [4], [5].

2.1.1 Species classification

Automated classification of bird sounds is a central task in bioacoustics and is widely used for wildlife monitoring. Current systems usually separate detection (finding vocalisations in recordings) from classification (identifying the species or call type). This two-step workflow improves accuracy by applying classification only to detected sounds and reducing the influence of large amounts of background audio [4], [6]. Although the term "classification" is sometimes used to include both steps, this thesis treats detection and classification as separate tasks.

Convolutional Neural Networks (CNNs) currently dominate this field, with widely used models including BirdNET, ConvNextBS, Perch, and SurfPerch [7]. Stowell proposed a practical guideline for deep-learning bioacoustics that recommends using well-known CNN architectures rather than designing custom networks [4].

Several variants have shown improved performance, such as CRNNs, which add an RNN layer after convolutional layers, and TNNs, which incorporate temporal modelling.

Many modern approaches use pre-trained models to generate embeddings, which are compact numerical representations of sound learned from large training datasets. These embeddings can then be used with a smaller, task-specific model trained on limited data. This allows similarity search, clustering, and detection of rare or unknown sounds. While common in other fields, embedding-based analysis is relatively new in passive acoustic monitoring and is strongly associated with large wildlife models such as BirdNET. Stowell also highlights few-shot learning, which enables models to learn new sound classes from only a small number of labelled examples, helping leverage diverse ecological datasets and improve generalisation [8]. He further notes increasing interest in multi-modal learning and possible future shifts toward Transformer-based models, although CNNs and spectrogram inputs

remain dominant, with raw-waveform and spectrogram-free approaches emerging as new directions [8].

2.1.2 Vocalisation classification

While bird acoustic classification often focuses on distinguishing species, another important task is identifying different call types within a single species. Although this area is still underexplored, it is the focus of our thesis, as we only study guillemots. Feature embeddings from pre-trained models like BirdNET can support this task. For example, they can separate adult and juvenile Great Gray Owls and identify multiple call types in Great Spotted Woodpeckers [9] as well as within-species classification of call types or life stages [10].
Within-species classification works best when relevant call types are included and enough examples are available. Rare signals can still be analysed using clustering or few-shot approaches. Pre-trained embeddings allow knowledge learned from large, species-level datasets to be applied to smaller, species-specific tasks, helping detect uncommon or previously unlabelled calls while supporting ecological and behavioural insights [10].
This approach lets passive acoustic monitoring capture both large-scale patterns and detailed information about behaviour, life stages, and population dynamics of vocal species.

2.1.3 Learning Paradigms and Model Architectures in Bioacoustic Sound Classification

Bioacoustic models use different learning paradigms depending on data availability. Supervised learning (SL) relies on labelled recordings and is used by CNN-based models like BirdNET, ConvNextBS, ViTINS, Perch, and SurfPerch. Self-supervised learning (SSL) uses unlabelled audio to learn representations via masked prediction, contrastive learning, or self-distillation (e.g., BirdMAE, (Bird)AVES, Animal2Vec, BioLingual), which can then be fine-tuned for few-shot or zero-shot tasks. This has enabled transformer models to model long-range temporal patterns and complex sequences in bird vocalisations [7].
Bioacoustic research often suffers from scarce labelled data. SSL transformers can learn birdsong structure autonomously, and unsupervised sound separation improves robustness to noise.
In 2021, Stowell highlighted SSL models like Microsoft BEATS, noting their advantages over other approaches were not yet clear [4].
By 2025, Ghani et al. compared convolutional neural networks (CNNs) and transformer-based models on large-scale birdsong datasets, evaluating BirdNET (CNN), PaSST (a spectrogram transformer with patchout), and PSLA (EfficientNet trained on AudioSet) across single- and multi-label classification tasks. Transfer learning proved highly effective, with bird-specific pre-training outperforming general audio pre-training. In particular, PaSST combined with knowledge distillation from BirdNET achieved the highest performance.
However, performance differences between models were relatively small compared to

differences in computational cost. While the transformer-based PaSST reached the highest mean average precision (mAP = 0.71), it required substantially longer training time (23.2 hours) compared to PSLA (5.21 hours) and BirdNET (0.59 hours), which achieved comparable performance (mAP = 0.67 and 0.68, respectively). This suggests that although transformers can offer slight gains in in-domain accuracy, CNN-based models remain competitive while being significantly more efficient.
Overall, CNNs demonstrated strong generalisation and efficiency, whereas transformers achieved marginally higher accuracy at a much higher computational cost. Additionally, shallow fine-tuning generalised best to new soundscapes, and knowledge distillation enabled flexible transfer across models [11].
Two complementary studies published in August 2025 examined large-scale pretrained bioacoustic models. Rauch et al. first introduced Bird-MAE, showing that domain-specific self-supervised pre-training combined with parameter-efficient probing approaches fine-tuning performance, improves frozen representations, and enables strong few-shot classification in bioacoustic tasks [12]. In parallel, Schwinger et al. evaluated foundation models and found that Bird-MAE achieved the best performance on BirdSet, while BEATsNLM performed slightly better on the broader BEANS dataset; transformer models benefited strongly from attentive probing, and supervised bird-specific CNNs such as Perch and ConvNextBS remained competitive in linear probing and resource-constrained settings [7].
Efforts are shifting toward more transparent and efficient models. CNNs remain widely used but are computationally heavy and difficult to interpret, while transformers and ConvNeXt provide higher performance yet still require substantial resources. Hopfield neural networks offer a lightweight and interpretable alternative that runs efficiently on standard laptops, adapts to different species or call types, and is suitable for monitoring rare species, although the evaluations from the paper we read are limited to simple binary classification of two bat species with distinct echolocation frequencies and lacked direct comparison with CNN-based approaches [13].

2.2 Multi-modal machine learning

Stowell noted the growing use of combining different input types (audio, video, text) by mapping each into a vector representation that can be merged or trained end-to-end [8].
Multi-modal systems often use two separate parts, one for audio and one for video/images. Each part turns its input into embeddings and the model then combines them to make a prediction. Common choices include spectrogram-based audio models and CNN/ViT-style vision models, with fusion done either by simple feature combining or by attention-based models that let the audio stream focus on useful visual features and the visual stream focus on useful audio features [14].
A common fusion strategy is late fusion, where each stream produces its own prediction score and the final decision is a weighted combination. For example, Tennekoon et al. classify bird behaviour using the image model InceptionResNetV2 on extracted frames and the audio model YAMNet embeddings with Logistic Regression, and combine them using weighted late fusion (0.4 video, 0.6 audio). They also

use YOLOv8 to focus on the bird in the frame and Grad-CAM to visualise which image regions drive the decision. In their setup, the fused model performs better than using audio or video alone [15].
One other approach is audio–language models, where audio is matched to text descriptions. Miao et al. show that CLAP can identify broader categories (e.g. whales, birds, frogs) using simple text prompts in a zero-shot setting [16]. They report weaker performance for fine-grained species labels, and results can vary with prompt wording, making prompt selection a practical limitation [16].
Combining audio and visual information can improve robustness when one stream is noisy or obstructed, but it increases computational cost and requires good time alignment between audio and video during fusion [14].

2.3 Input data

2.3.1 Common datasets

Many recent bioacoustic models are trained on large community datasets such as Xeno-Canto and related benchmark collections (e.g., BirdSet, BIRB, BEANS), which support model comparison and transfer learning [7]. BirdSet and BIRB are commonly used bird-focused benchmark subsets, while BEANS combines multiple datasets including birds and other animal groups. Schwinger et al. summarise which foundation models use which training splits and data sources (e.g., Xeno-Canto, Macaulay Library, and benchmark subsets) [7].

2.3.2 Input data type

Most bioacoustic deep learning systems use spectrograms as input, often with pseudo-log frequency scales such as mel or CQT. Stowell also notes that PCEN is often useful as spectrogram preprocessing [4]. Recent foundation models commonly use mel-spectrogram inputs and often apply logarithmic scaling and normalisation as part of preprocessing [7].
Different studies report different results for which spectrogram scaling works best (mel vs logarithmic vs linear), and some approaches stack multiple spectrogram representations of the same waveform as multi-channel input [4]. Another approach is to use the raw waveform as input instead of a spectrogram, this is supported by architectures such as WaveNet and temporal convolutional networks and but is often reported to require larger training datasets [4].
Since no feature representation is always best, it is worth comparing a few options in a small pilot study using a validation set. For noisy recordings with low SNR, it can help denoising before feature extraction [3]. Data augmentation is commonly used to increase variation in training data (e.g. time shifting, noise mixing, mixup-style combinations, time stretching/compression, or pitch shifting) [3], [4]. Kahl et al. found that mixup (mixing pairs of bird call recordings with random weights) and noise injection were especially beneficial for BirdNet's robustness [17]. Kershenbaum points out to be careful that augmentations should not invalidate labels [3].

Mel-frequency cepstral coefficients (MFCCs) have sometimes been used, but Stowell argues they are often a poor match for CNNs and are typically outperformed by less-preprocessed representations such as mel spectrograms [4].

2.4 Challenges in bioacoustics

Bioacoustic soundscapes often contain high levels of noise, strong class imbalance, and distribution shifts between training and deployment conditions [18]. Imbalance is commonly handled by using metrics such as macro-averaged scores, and adverse weather (wind and rain) remains challenging for automated analysis [4]. In out-of-domain environments, pretrained tools can also produce systematic errors; for example, BirdNET can falsely detect non-bird sounds [6].
In addition, overlapping calls from multiple individuals or species complicate detection, as many models assume a single dominant source. Modern bioacoustic datasets are also large and continuous, making storage and analysis computationally demanding. Finally, models must generalise across changing conditions while reducing reliance on manual annotation [19].

2.5 Studying guillemots

2.5.1 Guillemots vocalisations

Guillemots breed in dense cliff colonies where many individuals occupy fixed nest sites and interact repeatedly with neighbours. Although aggression toward unfamiliar birds is common, stable neighbour relationships are maintained within local groups.
Vocalisations are used in specific behavioural contexts such as territorial interactions, pair bonding, and parent–offspring coordination. Chicks begin producing simple peep calls before hatching, which later develop into more complex frequency-modulated calls after approximately two weeks.
Adults produce a small set of context-dependent calls, including crow (general interaction/aggression), laugh (pair and social bonding), nod/alarm (disturbance response), adow (copulation-related in females), and low-intensity growl calls [20].
However, evidence suggests that acoustic cues alone may not reliably encode individual identity during most of the breeding cycle, with visual familiarity and spatial location likely playing a major role in recognition. Acoustic signalling appears most critical in specific situations, particularly during chick departure from the colony and at-sea parent–chick contact, where maintaining communication is essential for feeding and survival [21].
In this thesis, we investigate whether guillemot vocalisations can be used for sound-only classification of behavioural events, specifically distinguishing between arrivals with a fish, arrivals without a fish, and periods with no arriving birds. In this context, guillemot vocalisations provide structured but highly context-dependent acoustic signals rather than continuous individual identifiers, with the strongest informative content expected during feeding-related interactions between returning

adults and chicks.

2.5.2 The AukLab

The Karlsö AukLab, established in 2008 with funding from WWF, is an artificial nesting platform for common guillemots (*Uria aalge*) located on the cliff edge of Stora Karlsö in the Baltic Sea. The structure extends approximately 30 m above sea level and consists of artificial rock ledges designed to resemble the natural habitat. It provides nesting sites for several hundred breeding pairs and allows continuous observation with minimal disturbance.

The ledge system consists of multiple artificial nesting sites embedded in the cliff face. These fixed locations are reused by the same individuals across breeding seasons, enabling consistent long-term behavioural monitoring.

2.5.2.1 Monitoring infrastructure and data acquisition

The AukLab supports continuous multi-modal data collection across the breeding season, including video, audio, thermal imaging, and environmental sensors.

The structure is organised into horizontal levels, numbered from 1 (bottom) to 9 (top), each containing multiple ledges (Figure 2.1). These ledges are grouped into vertical sections named after major guillemot colonies: Røst, Bonden, Farallon, Triangle, Björn, and Isle of Man. Ledge identifiers therefore encode both level and section (e.g., FAR3 refers to level 3 in the Farallon section).

In 2025, 14 ledges were monitored. The audio system consists of two components: (i) an 8-channel high-quality microphone array (DPA microphones), and (ii) 10 additional microphones integrated into the IP cameras.

The DPA microphones are connected to a Zoom F8 Pro field recorder, providing 8 synchronised audio channels. Each channel is assigned to a specific ledge:

- Ch1: BOND6
- Ch2: FAR3
- Ch3: TRI6
- Ch4: TRI7C
- Ch5: BOND1
- Ch6: ROST2
- Ch7: TRI2
- Ch8: Bjorn1

Four ledges are recorded by both DPA and camera microphones, allowing direct comparison between high- and low-quality audio and providing redundancy for validation and synchronisation.

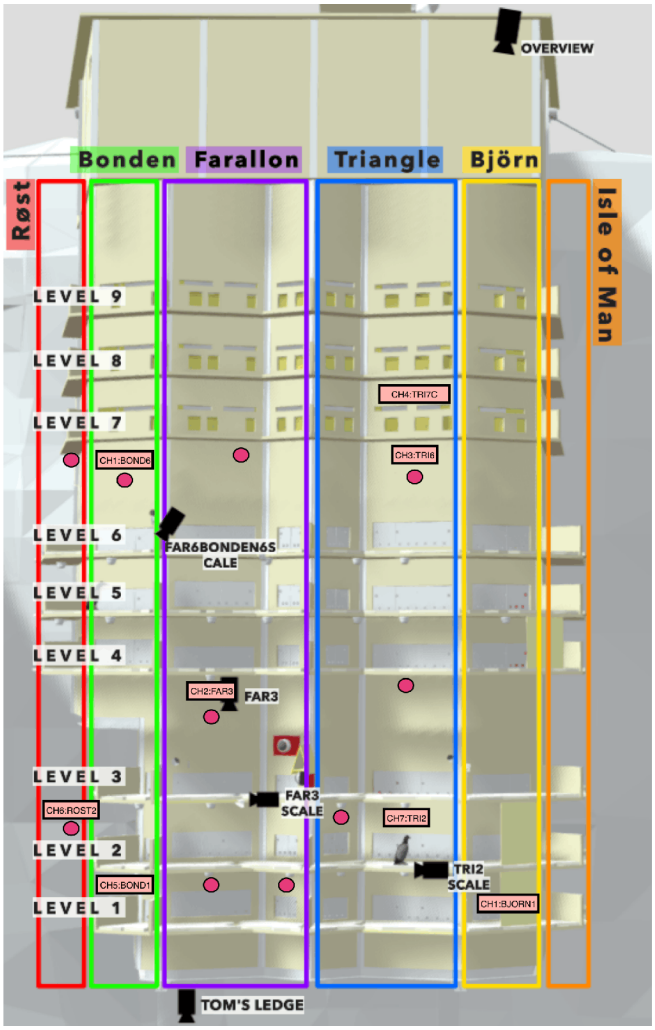

Figure 2.1: Ledge layout and placement of cameras and DPA microphones (red dots) in the 2025 setup.

The system produces approximately 16.9 TB of video data and 12 TB of audio data per season. Recordings are stored as ∼10-minute MKV (video) and RF64/WAV (audio) files.

2.5.2.2 System architecture and synchronisation

All data streams are synchronised using a GPS-based time system. A GPS pulse-per-second signal drives a dedicated Clock Pi, which acts as a Stratum-1 NTP server. All devices regularly synchronise to this clock. A backup NTP server provides redundancy.

The 8-channel audio from the Zoom F8 Pro is recorded by a Recording Pi and stored locally. An Analytics Pi monitors the recordings and transfers completed files to a NAS server. Transfers are verified using checksums, and system status is logged automatically.

2.5.2.3 Data management and autonomy

The system is designed for autonomous operation over long periods. Recording, transfer, and monitoring are fully automated, with built-in redundancy for both timing and storage. This ensures reliable, time-aligned data collection for large-scale multimodal analysis.

3

Methods

This study implements a machine listening pipeline to classify guillemot behavioural events from passive acoustic recordings, with the full workflow shown in Figure 3.1. The work is event-driven: YOLO detections from the colony camera footage are aggregated and smoothed to identify candidate arrival events, which are refined through confidence thresholding and manual video review to produce a ground truth labelled database on the FAR3 ledge. The audio from these events is extracted, pre-processed, and fed into each frozen pre-trained backbone models (BirdNET, PANN, and Perch), with a classifier head trained on the resulting embeddings to distinguish fish arrivals, no-fish arrivals, and non-events (background colony noise). Model performance is then evaluated through a suite of metrics and embedding analyses.

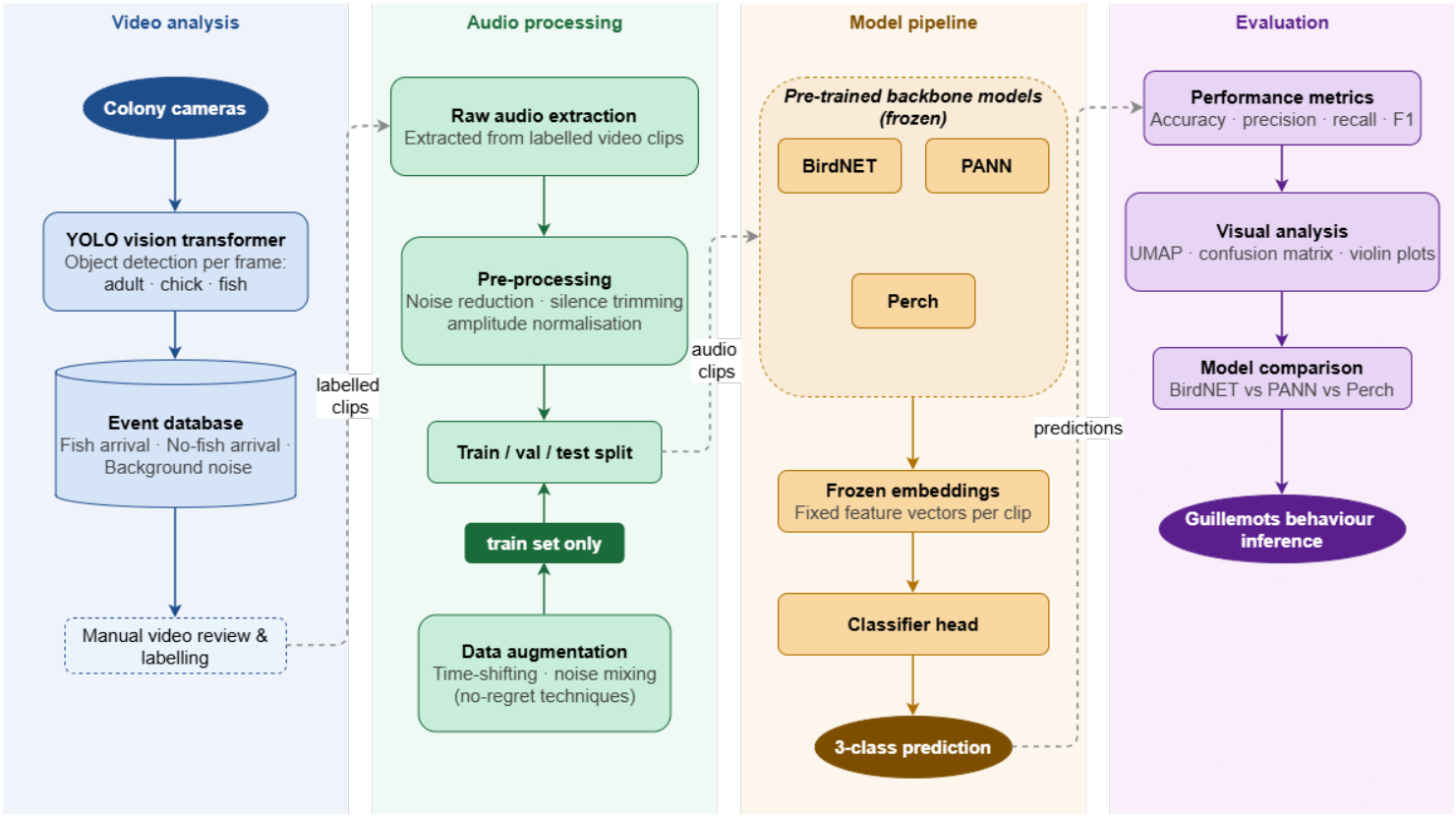

Figure 3.1: Method pipeline flowchart

3.1 Video analysis

YOLO (You Only Look Once) is a family of computer vision architectures designed for real-time object detection, classification, and localisation of objects within images or video frames in a single forward pass through a convolutional neural network (CNN). This single-stage design makes YOLO models particularly efficient in terms of inference speed while maintaining strong detection performance.
YOLOv26 is the most recent iteration in the Ultralytics YOLO series, released on 14 January 2026 [22].
In this study, object detection outputs were generated using the trained YOLOv26 model and provided by researchers at the AukLab. The detections were delivered as CSV files containing frame-level predictions for three classes: adult, chick, and fish. Given that the focus of this work is on guillemot arrival events in relation to fish presence, the dataset was restricted to camera sites with a high frequency of breeding pairs that successfully raised chicks, as well as locations where fish detection performance was assessed to be reliable.

3.1.1 Camera selection

3.1.1.1 Fish detection

To identify cameras where fish were reliably detected, we analysed the spatial distribution of detected fish by plotting the centre of each bounding box for every frame and camera.
A tight, well-defined cluster near the centre of the frame is a strong indicator that the model consistently detects fish at biologically meaningful locations, such as feeding areas. In contrast, detections that are uniformly distributed across the frame, or concentrated along image borders, are indicative of false positives. In these cases, the model is likely responding to background motion or frame edges rather than actual fish presence. In addition, the visualisation is colour-coded by detection confidence, where higher-confidence detections are shown in warmer colours (red) and lower-confidence detections in cooler colours (blue). This provides an additional layer of information for assessing detection reliability.
This approach allows us to filter the cameras without manual video review. For example, the FAR-3 camera shows several distinct and dense clusters near the centre of the frame, suggesting multiple stable detection zones and a strong correspondence with fish activity (Figure 3.2). In contrast, ROST2 displays clusters along the edges of the frame, which is typical of boundary-driven false detections rather than true fish occurrences (Figure 3.4). BONDEN6 shows intermediate behaviour, with more structured clustering than ROST2 but less central concentration than FAR-3 (Figure 3.3).
Only cameras equipped with high-quality DPA microphones were included in the analysis. These were: Ch1 (BOND6), Ch2 (FAR3), Ch3 (TRI6), Ch4 (TRI7C), Ch5 (BOND1), Ch6 (ROST2), Ch7 (TRI2), and Ch8 (Bjorn1).
Based on this visual inspection, FAR-3 and BONDEN-6 were selected as the most suitable sites for fish detection analysis. Among these, FAR-3 was chosen as the

primary focus for the study and used to construct the ground-truth database, as it demonstrated the most reliable and spatially consistent fish detection patterns.

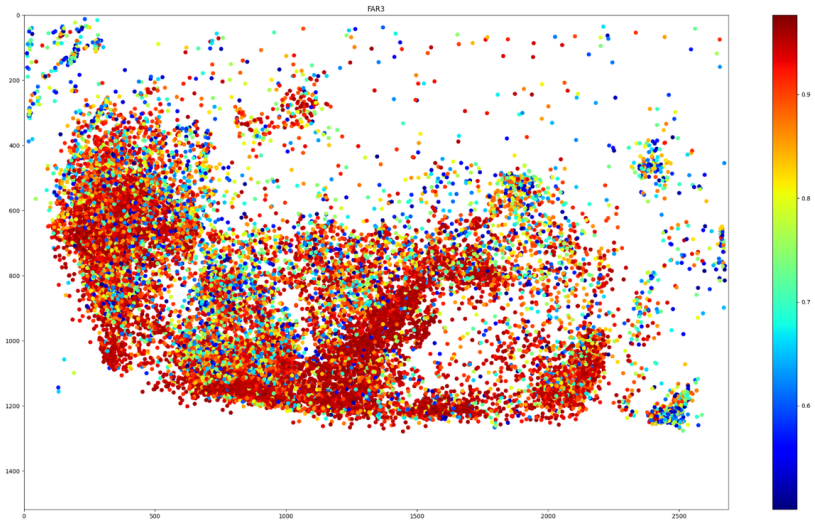

Figure 3.2: Spatial distribution of fish detection bounding box centres for FAR-3.

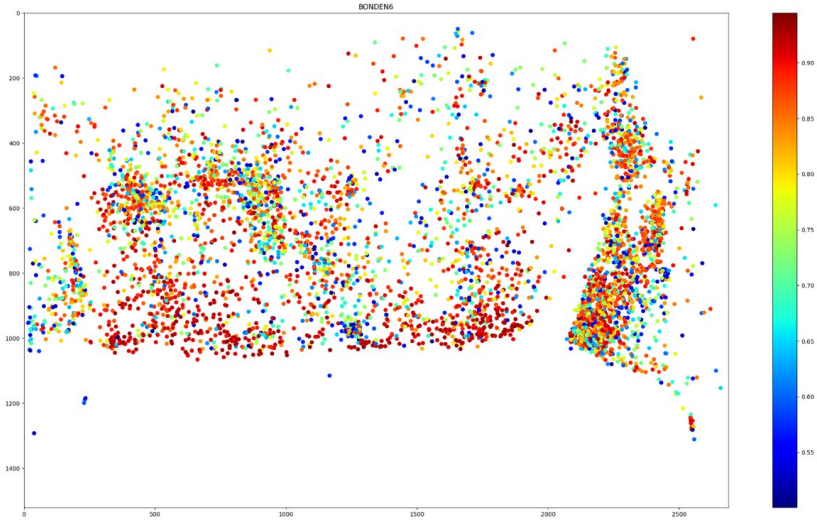

Figure 3.3: Spatial distribution of fish detection bounding box centres for BONDEN-6.

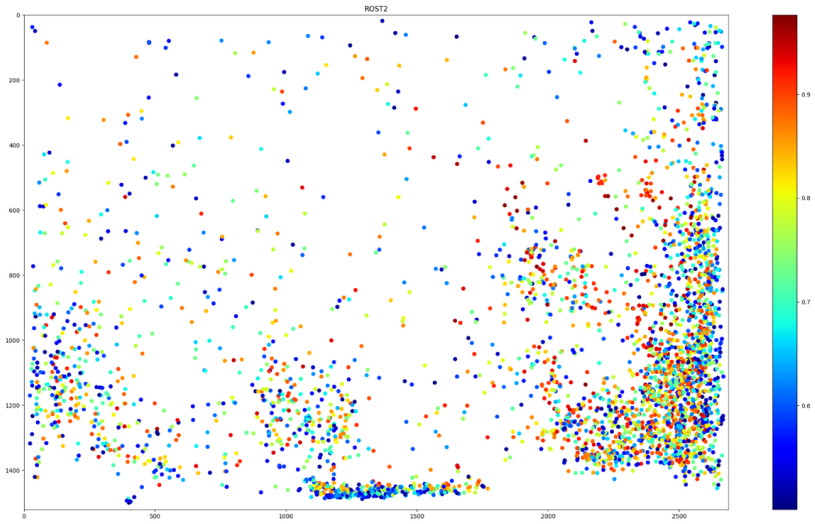

Figure 3.4: Spatial distribution of fish detection bounding box centres for ROST2.

3.1.1.2 Breeding summary of 2025

The next step was to examine the metadata associated with the FAR3 and BONDEN6 ledges during the 2025 breeding season. This information is essential for subsequent analyses, as it provides insight into how many breeding pairs occupied each ledge and how many of them successfully reproduced.
The AukLab database consists of the following tables:
`BreedingData, ChickIDPair, FeatherSamples, IndSex, PairIDYr, PairIndividuals, Recoveries, ResightingsAuklab, ResightingsNatural, Ringings, Ringvior, StationId`
Among these, *BreedingData* contains the most relevant information for this analysis, including the following variables:
`PairID, BrAttempt, EggDate, HatchDate, ChickGoneDate, EggLossDate, ChickLossDate, FailureCause, EggCol, BreedSucc, EggPosX, EggPosY, EggExact, HatchExact, FledgeExact, ChickMinAge, Comment, FailureType`
Querying the database for unique `PairID` values confirmed the presence of eight breeding pairs on the FAR-3 ledge. This was further validated through inspection of camera footage, which showed consistent occupancy patterns. Among these eight pairs, one lost its egg following a disturbance, likely due to intraspecific conflict.
For the BONDEN-6 ledge, five breeding pairs were identified, none of which lost their egg.
The timing of hatching and chick departure for FAR-3 and BONDEN-6 is summarised in Tables 3.1 and 3.2, respectively.
These datasets allow for filtering of relevant breeding events, as adult birds typically do not arrive with prey unless a chick is present. Based on this behaviour, the analysis was restricted to the period starting on 16 June, right before most hatching events occurred, and ending on 7 July, after which only a single chick remained on

Table 3.1: Breeding outcomes for FAR-3 pairs in 2025

PairID	HatchDate	ChickGoneDate	EggLossDate	ChickMinAge
FAR-3-2025-1	2025-06-13 21:58	2025-07-02 23:11	–	19
FAR-3-2025-2	2025-06-14 15:29	2025-07-07 21:23	–	23
FAR-3-2025-3	2025-06-17 07:59	2025-07-08 03:24	–	21
FAR-3-2025-4	2025-06-17 00:49	2025-07-06 22:52	–	20
FAR-3-2025-5	2025-06-17 00:42	2025-07-06 23:18	–	20
FAR-3-2025-6	2025-06-18 08:02	2025-07-07 21:52	–	20
FAR-3-2025-7	2025-06-17 19:47	2025-07-07 22:00	–	20
FAR-3-2025-8	–	–	2025-06-14 11:39	–

Table 3.2: Breeding outcomes for BONDEN-6 pairs in 2025

PairID	HatchDate	ChickGoneDate	EggLossDate	ChickMinAge
BONDEN-6-2025-1	2025-06-16 15:58	2025-07-07 20:01	–	21
BONDEN-6-2025-2	2025-06-18 23:19	2025-07-09 22:26	–	21
BONDEN-6-2025-3	2025-06-20 05:02	2025-07-07 22:25	–	18
BONDEN-6-2025-4	2025-06-20 06:04	2025-07-12 21:53	–	23
BONDEN-6-2025-5	2025-06-20 02:10	2025-07-07 22:27	–	18

the ledge. Additionally, extensive parent–chick vocal interactions during the fledging period were excluded, as they could introduce noise and confuse the machine learning model used for acoustic analysis.

3.1.2 Event-driven analysis

The YOLO object detection outputs were used to infer bird arrival events from video frames and to construct a labelled dataset for the machine listening classification pipeline.

Based on the temporal window defined in the previous section (16 June to 7 July 2025), this dataset was restricted to the corresponding period for the classification task.

Three event classes were defined:

- **Fish:** arrivals where an adult lands carrying a fish,
- **No fish:** arrivals where an adult lands without a fish,
- **Non-event:** periods with no arrival activity.

Non-events were included to ensure the model can distinguish arrival-related acoustic activity from background colony noise.

Adult and fish detections were aggregated into per-second counts and converted into continuous time series. The signals were then smoothed using a rolling mean (16-second window applied five times) to suppress short-term detection noise, and their temporal derivatives were computed to highlight changes in activity.

Fish arrival candidates were identified as time points where both adult and fish signals increased simultaneously. Events were separated using a threshold-based scan with a minimum inactivity gap of 10 seconds between consecutive events. Multiple

threshold values were evaluated against a manually curated dataset, and the optimal balance between false positives and false negatives determined the final threshold.

For each candidate event, the timestamp was refined by locating the first high-confidence fish detection in the database within a 60-second window preceding the smoothed signal peak, correcting for the lag introduced by the smoothing step. Each event was then stored as a fixed 30-second segment, consisting of 15 seconds before and 15 seconds after the refined timestamp.

Arrivals without fish were detected from increases in the adult signal only. Candidate events were filtered by requiring the adult signal to remain consistently above a locally estimated baseline. The event time was defined as the start of this sustained activity, and the same 30-second window was used.

Non-events were extracted from periods where both the adult and fish signals showed no activity for at least 30 consecutive seconds, ensuring the absence of arrival activity.

To address class imbalance, the no-fish and non-event classes were randomly downsampled to match the number of fish arrivals. Automatic construction of the dataset produced 465 samples per class after downsampling.

Finally, all fish arrivals were manually inspected to remove false positives, cases where the bird landed on the wrong ledge, and arrivals that occurred after the chicks had left the ledge. In total, 176 events were removed and 34 were added, reducing the fish arrivals from 465 to 323. No-fish arrivals were verified through manual review of a larger candidate pool, yielding 418 confirmed events. Non-events were matched to the number of no-fish arrivals, giving a final dataset of 1159 samples in total.

3.2 Audio Processing

3.2.1 Channel mapping validation

To verify that the metadata correctly maps cameras to audio channels, we extracted 30-second clips corresponding to manually identified fish arrival events on the FAR3 ledge. For each clip, we plotted waveforms and spectrograms across all channels and compared signal intensity.

As shown in Figure 3.5, the FAR-3 channel exhibits the strongest acoustic activity during the event, both in the waveform and spectrogram representations. This confirms that the correct mapping corresponds to channel 2.

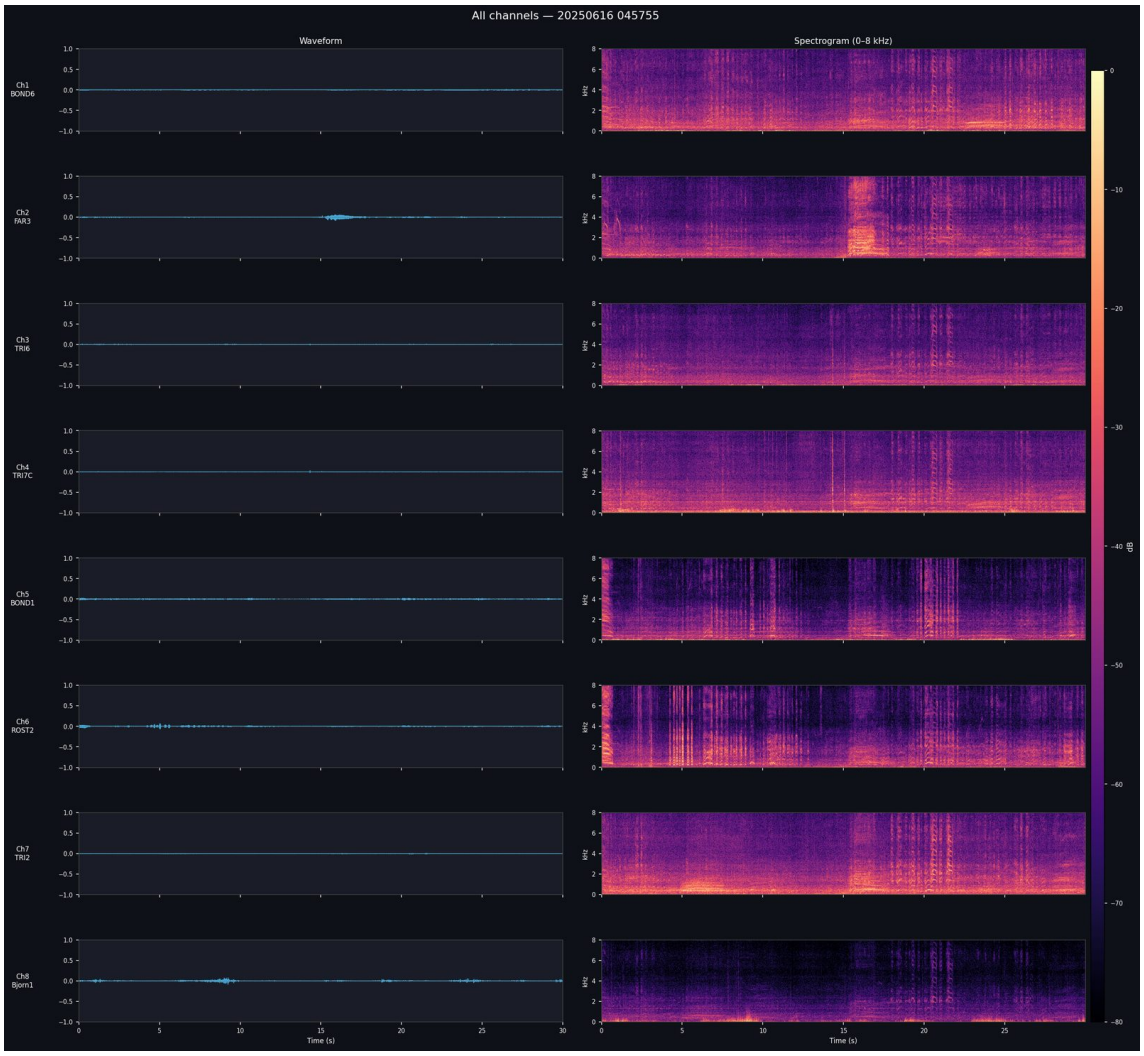

Figure 3.5: Comparison of spectrograms and waveforms across audio channels for a 30-second fish arrival event on FAR3.

3.2.2 Comparison of camera and DPA microphones

We compared audio from the camera microphones and the dedicated DPA microphones using identical 30-second segments extracted from fish arrival events.

The DPA recordings show substantially lower amplitude due to intentionally conservative gain settings. This prevents clipping during close-range vocalisations. In contrast, the camera microphones exhibit frequent saturation at the digital limit, indicating clipping during high-intensity events.

Since recordings are made at 32-bit depth, low recorded amplitudes are not problematic in this case. The increased dynamic range allows the signal to be safely scaled during post-processing without loss of information or introduction of clipping.

To make the signals comparable, the DPA audio was scaled to match the RMS level of the camera recordings. After scaling, the DPA signal reveals a cleaner waveform with preserved structure, while the camera signal shows distortion consistent with clipping.

For this part of the analysis (comparing the microphones), no additional preprocessing was applied, as the goal was to directly compare raw recording characteristics and their effect on downstream analysis.
Figure 3.6 illustrates the waveform comparison between the two microphone systems for the same event.

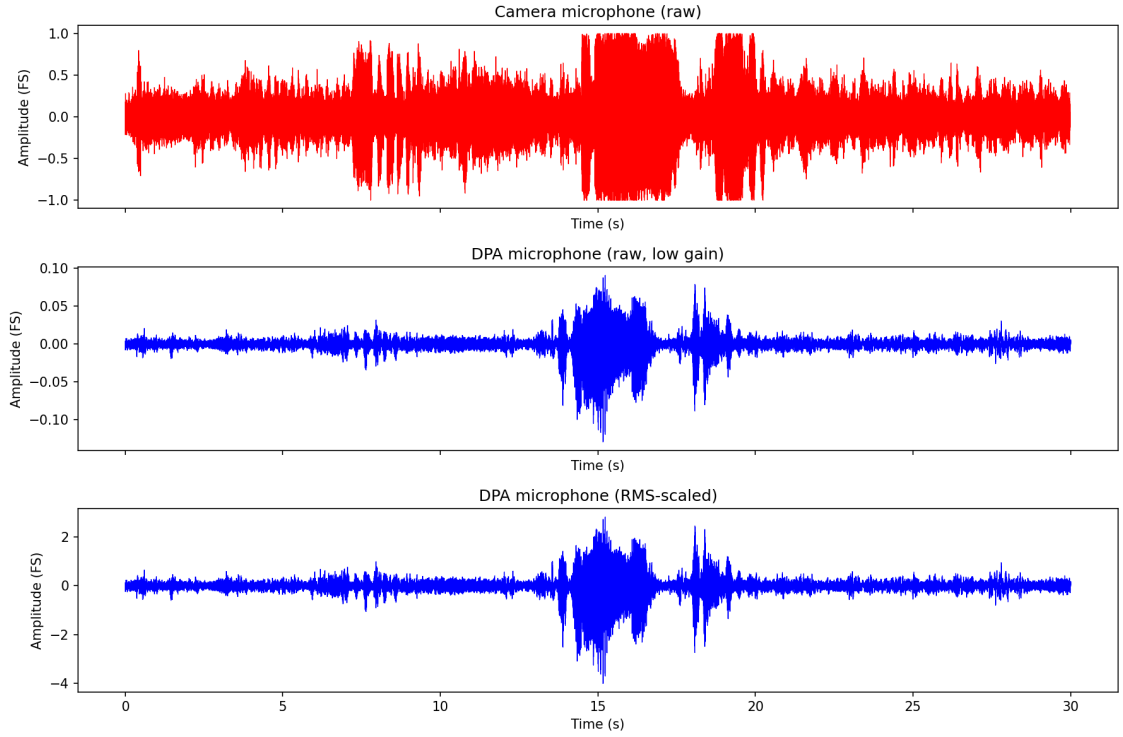

Figure 3.6: Waveform comparison between camera and DPA microphones for the same fish arrival event. The DPA signal is recorded at lower gain but, after RMS scaling, shows a cleaner representation. The camera signal exhibits clipping during peak activity.

3.2.3 Pre-processing

Audio was extracted from 10-minute multichannel WAV recordings using FFmpeg. A single microphone channel was selected (channel 2 for FAR3). Each clip is centred on the event midpoint, extending 15 seconds before and after, giving a fixed 30-second window. Where a clip crosses a file boundary, segments from two consecutive recordings are concatenated. All clips are mono and sampled at 48 kHz.
Clips were then preprocessed in two steps: (1) a bandpass filter (100–12000 Hz) removes noise outside the frequency range of relevant guillemot vocalisations; (2) RMS normalisation to a target level of 0.1 ensures consistent amplitude across recordings. Several filter ranges were evaluated, including a narrower 500–6000 Hz band more closely aligned with guillemot call frequencies; the broader 100–12000 Hz range was ultimately selected as it led to the best classification performance.

3.2.4 Data Augmentation

Augmentation was applied to the training set to improve model robustness. We used only "no-regret" techniques that preserve the acoustic content of a call [4]:

- **Time-shifting:** the waveform is circularly shifted by a random offset in $[-T/4, T/4]$.
- **Noise injection:** Gaussian noise is added with standard deviation in $[0.0005, 0.002]$.
- **Gain perturbation:** amplitude is scaled by a random factor in $[0.9, 1.1]$.
- **Mixup:** two training samples from the same class are blended as $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$, with $\lambda \sim \text{Beta}(0.2, 0.2)$, producing $\lfloor N/2 \rfloor$ additional samples per class.

Each augmentation is applied independently with probability 0.5. Time stretching and pitch shifting were tested but excluded, as even small changes ($\pm 3\%$ stretch, ± 0.5 semitones) produced artificial clustering in UMAP projections of BirdNET embeddings: augmented clips formed isolated clusters detached from the main distribution, indicating that the model captured transformation artefacts rather than biologically meaningful features. Each training clip was augmented twice, giving a training set roughly three times its original size.

3.2.5 Data Split

Files were split into training (80%), validation (10%), and test (10%) sets at the recording level, before augmentation, using a fixed random seed (42). Augmented samples therefore appear only in the training set. Validation and test sets received preprocessing only. The validation set was used for hyperparameter tuning; the test set was used once to evaluate the final model.

3.3 Models pipeline

Three frozen pre-trained CNN-based models were evaluated as feature extractors for the classification task: BirdNET, PANN, and Perch. Each model processes audio through overlapping sliding windows, producing a fixed dimensional embedding for each window. These embeddings are then pooled (see below) into a single vector representing the full 30 second event, which is passed through a classifier trained to distinguish the three classes (arrival with fish, arrival without fish, non-event).

3.3.1 BirdNET

BirdNET is a bird sound classifier based on a residual network (ResNet) architecture with 157 layers and over 27 million parameters, trained on large scale ornithological data with domain-specific augmentation [17]. It is designed to be robust to high ambient noise and overlapping vocalisations. BirdNET processes audio internally in approximately 3-second segments using the `birdnetlib` interface, producing one 1024-dimensional embedding per segment.

3.3.2 PANN

PANNs (Pretrained Audio Neural Networks) [23] is a class of large-scale audio neural networks trained on AudioSet, a large-scale audio dataset with an ontology of 527 sound classes consisting of over two million audio clips sourced from YouTube. The PANN CNN14 variant, a 14-layer convolutional neural network, was used here as a feature extractor to produce fixed-length embedding features for audio clips. Each audio file was sliced into overlapping 10-second windows with a 5-second hop, producing one 2048-dimensional embedding feature vector per window via the `panns_inference` AudioTagging interface.

3.3.3 Perch

Perch [24] (`perch_v2`) is a bioacoustic foundation model developed by Google DeepMind, based on an EfficientNet-B3 convolutional residual network with 12 million parameters. It was trained in supervised fashion on a multi-taxa dataset combining Xeno-Canto, iNaturalist, and other repositories, covering over 1.5 million recordings across 14,795 species and sound classes.
Audio was divided into non-overlapping 5-second windows, each producing a 1536-dimensional mean embedding through the `perch_hoplite` interface.

3.3.4 Window pooling

Because each 30-second event produces multiple window-level embeddings, a pooling step was required to obtain a single fixed-size representation per event. For each file, the window embeddings were stacked into a matrix $\mathbf{W} \in \mathbb{R}^{T \times D}$, where T is the number of windows and D is the window embedding dimension. The per-file embedding was then computed as the concatenation of the mean and standard deviation across the time axis:

$$\mathbf{f} = [\, \bar{\mathbf{w}} \parallel \sigma(\mathbf{W}) \,] \in \mathbb{R}^{2D} \tag{3.1}$$

where $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t$ and $\sigma(\mathbf{W})$ is the standard deviation across windows. This gives final per-file embedding dimensions of 2048 (BirdNET), 4096 (PANN), and 3072 (Perch). Max pooling was also evaluated, which selects the embedding of the highest-scoring window rather than aggregating all windows, as advised by Ghani et al. [11], but mean+std pooling performed best overall.

3.3.5 Classifier head

A lightweight classifier was trained on top of each pre-trained model's pooled embeddings, consisting of a single hidden layer with 256 units, ReLU activation, followed by a linear output layer over the three target classes. Training used Stochastic gradient descent (SGD) with momentum 0.9, a learning rate of 0.001, dropout of 0.3, and cross-entropy loss over 50 epochs with a batch size of 32. The learning rate was halved when validation accuracy showed no improvement for 5 consecutive epochs. Logistic regression and SVM classifiers were also evaluated as lightweight baselines using scikit-learn defaults.

3.4 Evaluation

3.4.1 Model Performance Comparison

All three backbone models (BirdNET, PANN, Perch) were evaluated on the same held-out test set using accuracy, per-class precision, recall, and F1-score. Training was performed on a balanced dataset, while inference was run on the natural class distribution to reflect realistic deployment conditions.
Accuracy measures the proportion of correctly classified samples across all classes:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

Precision quantifies how many predicted instances of a class are correct, and recall measures how many true instances of a class are correctly identified:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

The F1-score, defined as the harmonic mean of precision and recall, provides a single aggregated measure of class-wise performance:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Overall performance is reported using macro-averaged precision, recall, and F1-score, treating each class equally regardless of frequency:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c \tag{3.5}$$

where C is the number of classes and F1_c is the F1-score for class c.
Five visualisations were produced for each model. UMAP projections of the embeddings were computed after PCA pre-processing (retaining 90% of variance) to visualise class separability in two dimensions. Confusion matrices were plotted to show where each model confuses classes. Per-class F1 scores were compared across models as a bar chart. Confidence distributions for correctly labelled samples were shown as violin plots, revealing how decisively each model separates the classes. In addition to the MLP classifier head, logistic regression and SVM classifiers were evaluated. All three classifier heads were trained on the same frozen embeddings and evaluated on the same test split, with performance reported via classification reports (per-class precision, recall, and F1) and confusion matrices.

3.4.2 BirdNET Embedding Inspection

To understand what acoustic patterns drive BirdNET predictions, four analyses were performed on the test set using a logistic regression probe fitted on the training split pooled file-level embeddings.

Top confident recordings (30-second files). Each test recording was assigned a confidence score $P(\text{class} \mid \text{file})$ by the logistic regression probe, and the top 10 recordings per class were saved as spectrograms and audio files for manual inspection.
Top confident windows (3-second clips). Since the logistic regression classifier operates on pooled file-level embeddings, window-level inspection is used only for qualitative analysis. For each class, the top 10 recordings ranked by $P(\text{class} \mid \text{file})$ are selected, and within each recording the 3-second window with the highest BirdNET embedding norm is extracted. This yields one representative segment per file, highlighting the most acoustically salient regions within recordings that the model is already highly confident about.
Misclassified recordings. Recordings where the probe predicted the wrong class with confidence ≥ 0.80 were identified, highlighting potential labelling ambiguities or acoustic overlap between classes.
Average spectrograms and difference maps. Mean mel spectrograms were computed per class from original (non-augmented) test recordings. Pairwise differences $\bar{S}_A - \bar{S}_B$ were then plotted, with red regions indicating more energy in class A and blue in class B, revealing which frequency bands and time positions best separate each pair of classes.
All outputs are discussed in the results section.

3.4.3 Weather data

To assess whether environmental data improve classification performance, weather inputs were incorporated as additional features alongside the BirdNET embeddings. Weather data were retrieved from a SQLite database provided by the researchers at Stora Karlsö and joined to each audio recording by nearest-neighbour matching in time, with a tolerance of 15 minutes. Recordings that could not be matched were substituted using column medians. To avoid introducing redundant or misleading features, a stability-based feature selection step was performed before augmentation. A logistic regression model with elastic net regularisation was fitted repeatedly on bootstrap samples of the training data, and each feature was scored by the fraction of iterations in which it was selected among the top 25% most influential variables. The four most stable features were kept. The selected weather features were standardised using a scaler fitted exclusively on the training set and then applied to the validation and test sets, preventing any information leakage. The standardised features were concatenated to the pooled BirdNET file-level embeddings, producing an augmented representation that was passed to the classifier head instead of the acoustic embeddings alone.

4

Results and Discussion

This chapter presents and discusses the results of the classification pipeline. Section 4.1 compares the three frozen backbone models (BirdNET, PANN and Perch) on the manually curated FAR3 dataset, evaluating both the structure of their embedding spaces and their classification performance. Section 4.2 focuses on BirdNET for a more detailed analysis, covering the effect of the classifier head, the acoustic characteristics that drive its predictions, its misclassification patterns, the impact of incorporating weather data, and its ability to generalise beyond the ground-truth dataset and across ledges.

4.1 Backbone Model Comparison *(manual FAR3 dataset)*

The ground truth dataset was constructed from the FAR3 ledge, covering the period 16 June to 7 July 2025. After automatic event extraction and manual inspection, the final dataset consisted of 323 fish arrivals, 418 no-fish arrivals and 418 non-events, giving 1159 samples in total.

4.1.1 Class separability in embedding space

Before evaluating the performance of the MLP classifier, the structure of the learned embeddings is examined using UMAP projections of the training set. UMAP (Uniform Manifold Approximation and Projection) is a non-linear dimensionality reduction technique that preserves local neighbourhood structure, allowing high-dimensional embeddings to be visualised in two dimensions while retaining meaningful cluster relationships.

The embeddings produced by BirdNET, PANN, and Perch on the training subset are compared using these UMAP projections to assess how well each model separates the three acoustic classes in the embedding space (Figure 4.1).

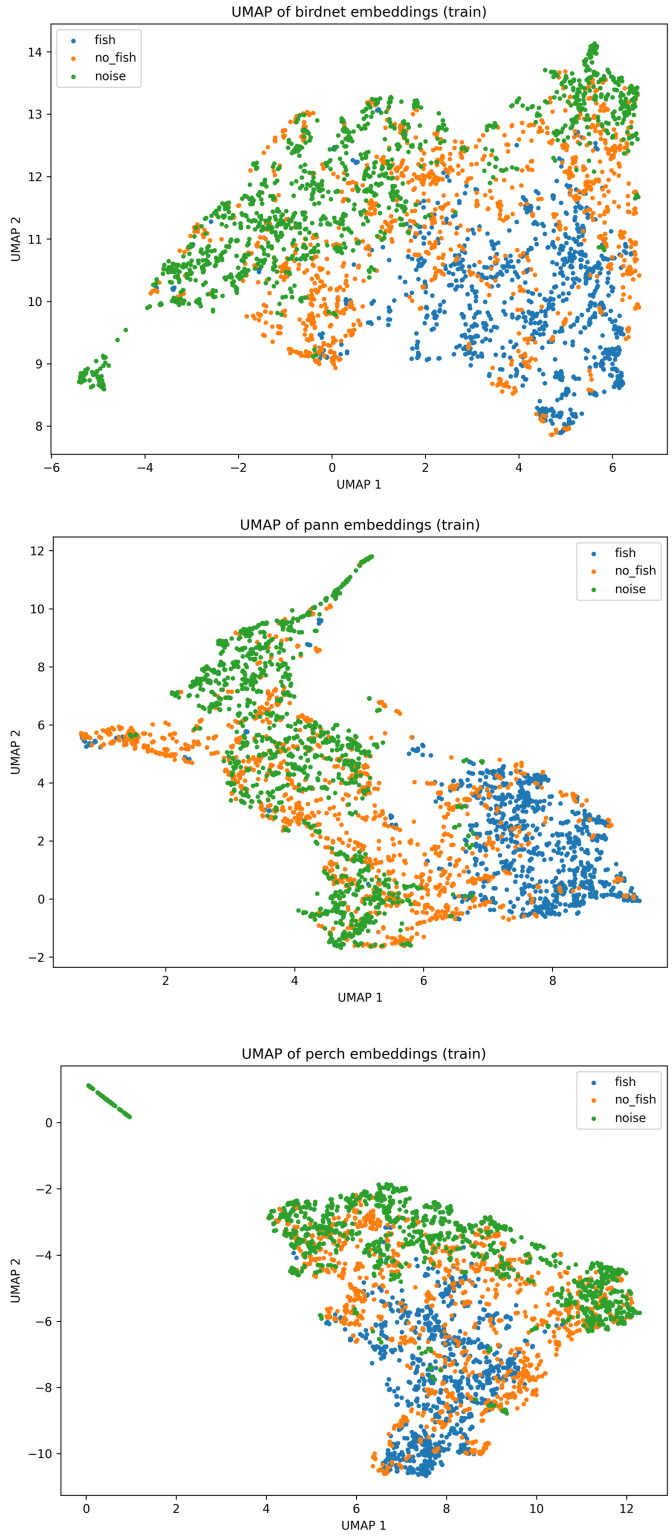

Figure 4.1: UMAP projections of embeddings for BirdNET (top), PANN (middle), and Perch (bottom).

BirdNET shows the clearest class separation of the three models. Fish (blue) forms a relatively compact region in the bottom right of the plot. Fish arrivals appear as

the most acoustically distinct events, pulling away into their own region, while no-fish shares characteristics with both fish arrivals and non-events, consistent with the idea that a quiet arrival with colony response can resemble either class depending on its intensity. A small subset of non-event points is also isolated in the bottom left, away from the main cluster, suggesting a distinct type of background recording that is acoustically different from typical colony noise.

PANN exhibits a similar overall organisation, with fish mostly concentrated on the right and non-events in the upper left. However, class boundaries are less distinct, with stronger overlap between no-fish and non-events. No-fish also forms a small cluster on the left side of the embedding space, indicating a subset of recordings represented differently from the rest. As with BirdNET, a subset of non-event samples scatters away from the main cluster, suggesting that background acoustic conditions vary rather than conforming to a single, compact noise type.

Perch shows weaker separation between classes. All three classes largely occupy a single dense region with significant overlap, indicating limited class structure in the embedding space. A small isolated streak of non-event points also appears in the upper left, which may reflect the same atypical colony noise or artefacts.

Overall, BirdNET shows the clearest class structure, with PANN displaying a similar but less distinct separation. Perch shows the weakest separation, with strong overlap between fish and no-fish, although non-event remains somewhat more distinct.

4.1.2 Classification performance of BirdNET, PANN and Perch

The three pre-trained backbone models (BirdNET, PANN, and Perch) were evaluated on the held-out test set using frozen embeddings and a lightweight MLP (Multi Layer Perceptron) classifier head. Table 4.1 summarises overall performance and Table 4.2 presents per-class precision and recall.

Model	Loss	Accuracy	Precision	Recall	F1
PANN	0.4949	0.7815	0.7875	0.7914	0.7894
PERCH	0.7068	0.6471	0.6378	0.6533	0.6455
BIRDNET	0.5321	**0.8067**	**0.8163**	**0.8170**	**0.8167**

Table 4.1: Overall test performance

Model	Class	Precision	Recall	F1-score
PANN	fish	0.8824	0.9091	0.8955
	no_fish	0.7073	0.6744	0.6905
	non-event	0.7727	0.7907	0.7816
PERCH	fish	0.6857	0.7273	0.7059
	no_fish	0.5484	0.3953	0.4595
	non-event	0.6792	0.8372	0.7500
BIRDNET	fish	0.8857	0.9394	0.9118
	no_fish	0.8182	0.6279	0.7105
	non-event	0.7451	0.8837	0.8085

Table 4.2: Per-class precision, recall, and F1-score

BirdNET achieved the strongest overall performance, with an accuracy of 80.7% and balanced precision and recall across all three classes. Its advantage was most noticeable for fish arrivals, where it reached a recall of 93.9%, meaning it correctly identified the large majority of true fish arrival events. This is particularly desirable in the current ecological monitoring context, where missed fish deliveries would directly undercount chick provisioning rates. The no-fish class proved the most challenging for BirdNET, with recall dropping to 62.8%, suggesting that no-fish arrivals share acoustic characteristics with both fish arrivals and non-events, making them harder to separate. PANN performed competitively, achieving 78.2% accuracy and strong fish detection (recall 90.9%), though it fell short of BirdNET across all aggregate metrics. PANN showed more balanced per-class performance than BirdNET, with less pronounced degradation on the no-fish class. This suggests that AudioSet-trained general audio features capture some relevant acoustic variation, though they are less specialised than BirdNET's bioacoustic representations for this task. Perch performed considerably worse than both other models, with an accuracy of 64.7% and a particularly low recall for the no-fish class (39.5%). Despite being trained on a large multi-taxa bioacoustic dataset, Perch's embeddings appear less discriminative for the specific vocalisations and colony soundscapes present at Stora Karlsö. Its relatively high non-event recall (83.7%) suggests it defaults toward predicting non-event when uncertain, at the cost of missing true arrival events. Across all three models, the no-fish class consistently showed the weakest performance. This is ecologically interpretable: arrivals without fish tend to be acoustically quieter, with fewer chick peep calls and a more subdued adult vocalisation, making them difficult to distinguish from periods of general colony activity.

The row-normalised confusion matrices in Figure 4.2 show the per-class error patterns for each model. BirdNET misclassifies 6% of fish arrivals as no-fish and none as non-event, and confuses 30% of no-fish arrivals with non-events. PANN shows a similar pattern, with 9% of fish misclassified as no-fish and 23% of no-fish misclassified as non-events. Perch shows the weakest separation, with 27% of fish misclassified as no-fish and no-fish split almost evenly between the three classes (40% correct, 21% predicted fish, 40% predicted non-event).

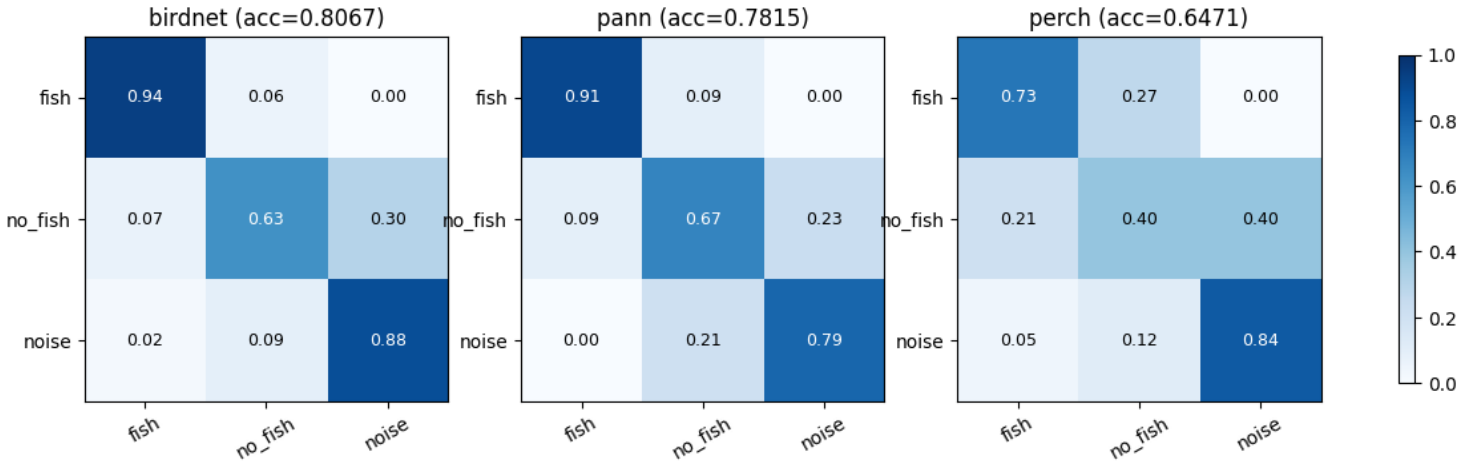

Figure 4.2: Row-normalised confusion matrices for BirdNET, PANN, and Perch on the test set.

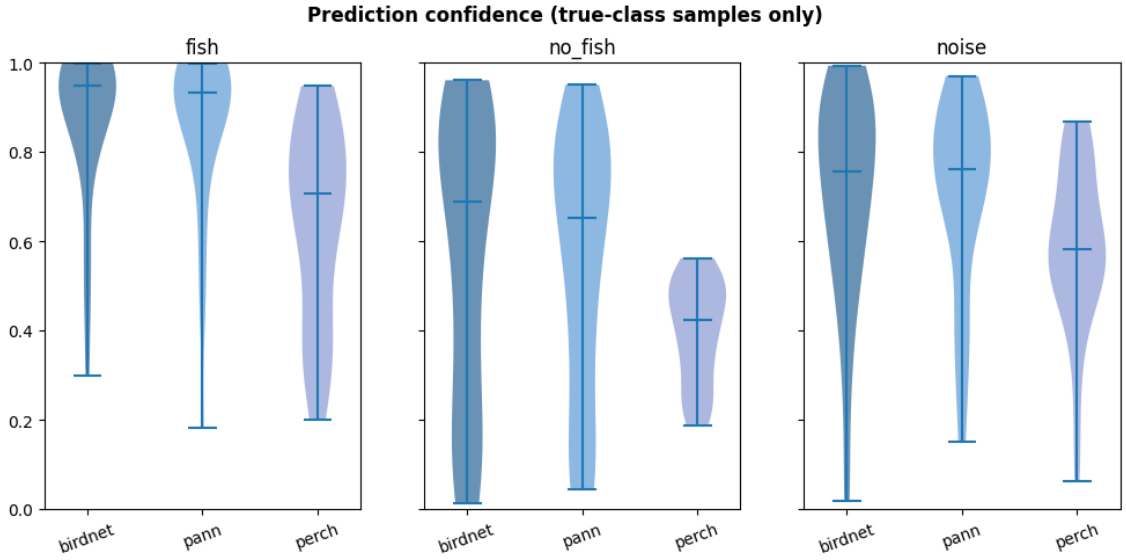

Figure 4.3: Prediction confidence distributions on true-class test samples for BirdNET, PANN, and Perch, shown per class.

Prediction confidence distributions for true-class samples are shown in Figure 4.3. For fish arrivals, BirdNET and PANN show wide distributions concentrated near 1.0, indicating that when correctly classified, fish arrivals are predicted with high certainty which is consistent with fish being the most acoustically distinct class. Perch shows a broader distribution at lower confidence values, suggesting its embeddings capture less of the information that distinguishes fish arrivals from the other classes. For no-fish arrivals, BirdNET and PANN show wide distributions spanning the full confidence range: even correct no-fish predictions are often made at low confidence, reflecting the genuine acoustic ambiguity of the class rather than a systematic modelling failure. Perch is narrowly concentrated at low confidence values, barely above chance. For non-events, BirdNET and PANN again show wide

distributions, and Perch shows a narrower distribution at moderate confidence values, consistent with its tendency to default towards non-events when uncertain, as seen in its high non-event recall elsewhere.

4.2 BirdNET in-depth Analysis *(Manual FAR3 dataset)*

Having established BirdNET as the best-performing model, this section examines it in greater depth. It uses the manually curated FAR3 dataset as ground truth and a starting point for comparison. It investigates how the choice of classifier head affects performance, which acoustic features drive its predictions, where it tends to fail, and how well it generalises beyond the ground-truth dataset. In doing so, we also use the model to better understand guillemot vocalisations.

4.2.1 Effects of Classifier Head on Performance

We evaluated the BirdNET embeddings using three different classifier heads: a multilayer perceptron (MLP), a support vector machine (SVM), and logistic regression. Hyperparameter tuning was not performed, as preliminary experiments showed no meaningful improvement on the test set. Given the computational cost, we prioritised model evaluation for downstream biological analysis rather than exhaustive classifier optimisation.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.80	0.79	0.80	0.80
SVM	0.80	0.80	0.80	0.79
MLP	**0.81**	**0.82**	**0.82**	**0.82**

Table 4.3: Overall performance of classifier heads on BirdNET embeddings.

Model	Class	Precision	Recall	F1-score
Logistic Regression	Fish	0.89	0.97	**0.93**
	No fish	0.75	0.70	**0.72**
	Non-event	0.77	0.77	0.77
SVM	Fish	0.89	0.94	0.91
	No fish	0.78	0.65	0.71
	Non-event	0.75	0.84	0.79
MLP	Fish	0.89	0.94	0.91
	No fish	0.82	0.63	0.71
	Non-event	0.75	0.88	**0.80**

Table 4.4: Per-class performance for BirdNET embeddings with different classifier heads. Best F1-score per class is highlighted in bold.

All three classifier heads achieved comparable overall performance, with accuracy and F1-scores within one percentage point of each other. This suggests that the discriminative information is largely captured by the BirdNET embeddings themselves, and that the choice of classifier head has limited impact on overall performance. The most visible differences appear at the per-class level. Logistic regression achieved the highest fish recall at 0.97, meaning it missed very few true fish arrival events, which is the most ecologically critical class. While the MLP achieved slightly higher no-fish precision (0.82 vs 0.75) and the best non-event F1-score (0.80), it struggled most with no-fish recall (0.63), meaning it more frequently failed to detect quiet arrivals altogether. The SVM showed a similar pattern, with no-fish recall of 0.65, though it achieved more balanced non-event performance. Across all three classifiers, the no-fish class consistently showed the weakest performance, reinforcing the finding that arrivals without fish are acoustically the most ambiguous class. Given the marginal differences in overall performance, logistic regression is preferred for its interpretability and its consistently stronger recall across both event classes, directly minimising missed prey delivery events in ecological monitoring.

4.2.2 Acoustic Characteristics of BirdNET Embeddings

4.2.2.1 Average spectrograms and difference maps

Figure 4.4 shows the mean mel spectrogram for each class in the test set. The fish class has a clear broadband burst around second 15, matching the moment of arrival. The no-fish and non-event classes do not show this pattern; both are mainly dominated by low-frequency energy across the full 30-second window. However, no-fish has slightly more mid-frequency activity than non-event and shows a faint trace of the broadband pattern seen in fish.

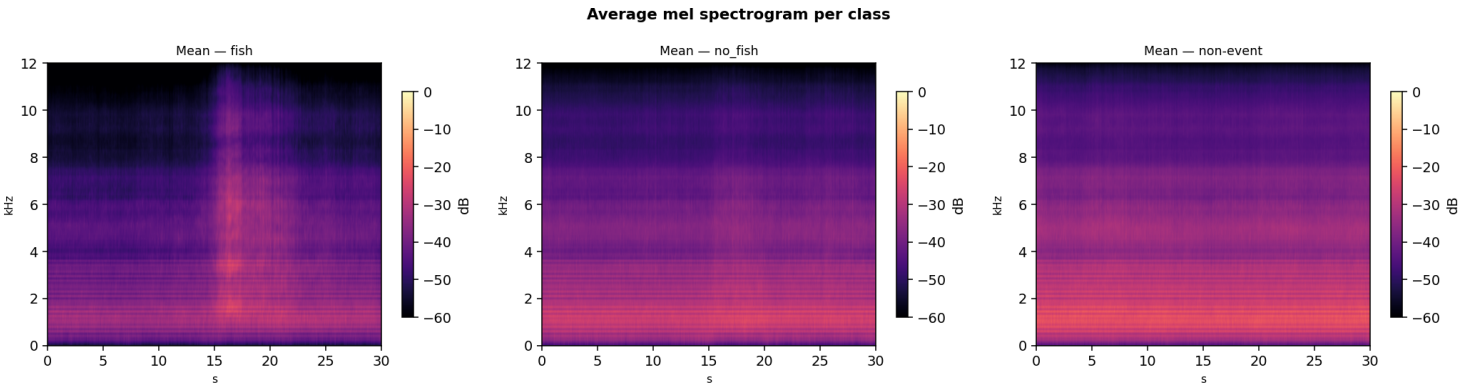

Figure 4.4: Mean mel spectrogram per class computed from original test recordings

The pairwise difference maps shows a clearer picture of how each pair of classes can be separated in the spectral domain. It is important to note that the colour scales differ across maps: the fish versus non-event map spans ±17 dB, fish versus no-fish spans ±15 dB, and no-fish versus non-event spans only ±10 dB. This alone shows that fish arrivals are easier to separate from the other two classes than no-fish and non-event are from each other.

In the fish versus no-fish map (Figure 4.5) and the fish versus non-event map (Figure 4.6), the pattern is clear: a strong red vertical column at second 15 confirms that the arrival burst is the primary feature separating fish from both other classes. Everything outside that window is blue, meaning the other classes carry more low-frequency energy during quiet moments. The fish versus non-event contrast is larger in magnitude, consistent with non-event recordings having a higher background noise than no-fish recordings.

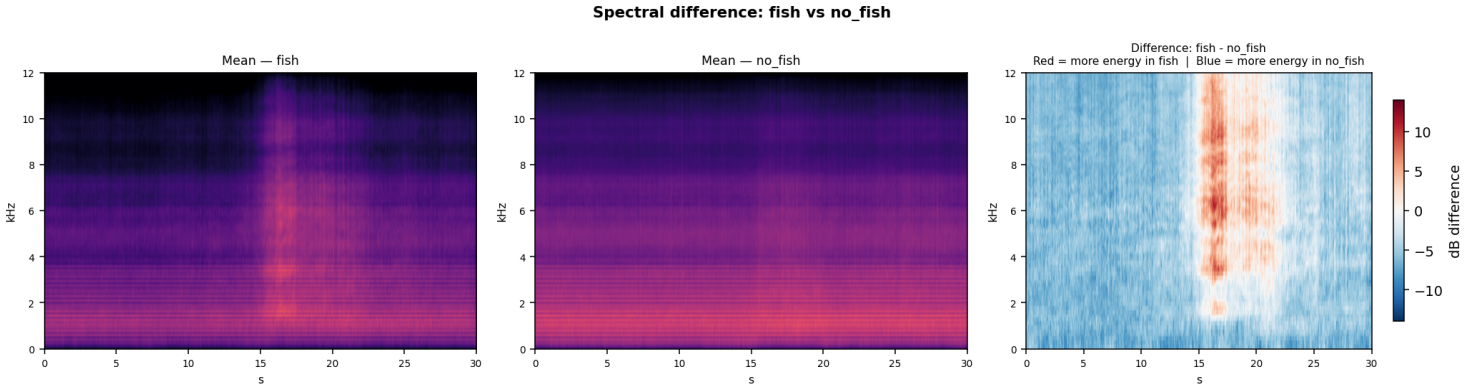

Figure 4.5: Pairwise spectral difference between fish and no-fish classes (±15 dB scale).

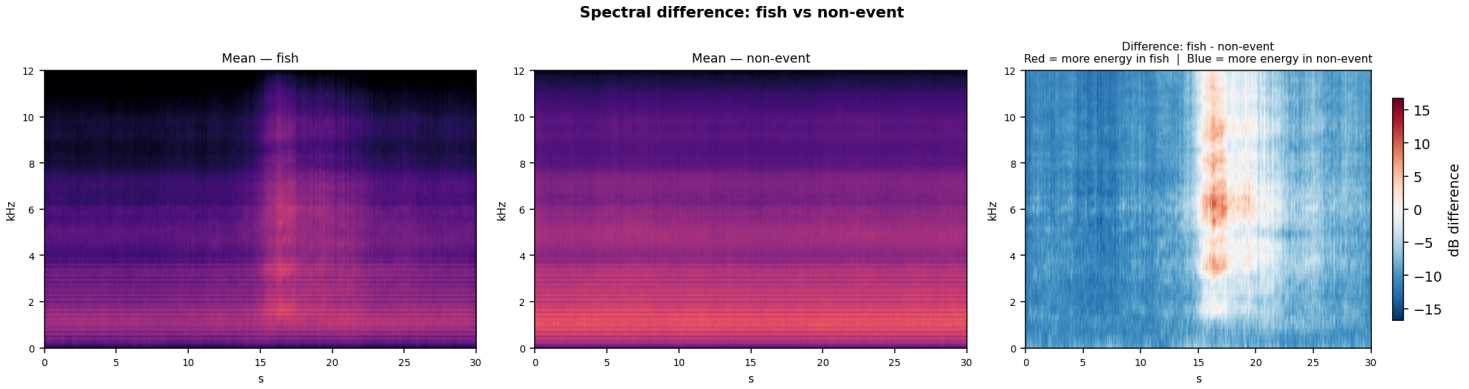

Figure 4.6: Pairwise spectral difference between fish and non-event classes (±17 dB scale).

The no-fish versus non-event map (Figure 4.7) is the most revealing for understanding classifier difficulty. The scale is only ±7 dB and the map is almost uniformly blue, confirming that these two classes are acoustically very similar. However, a faint white vertical band is visible around second 15, with small red sparks in the 2–6 kHz range. This suggests that no-fish contains a weak but consistent increase in energy at the expected arrival time compared to noise. As a result, the model is not relying on random variation, but on a subtle signal, which helps explain why no-fish is still detectable above chance.

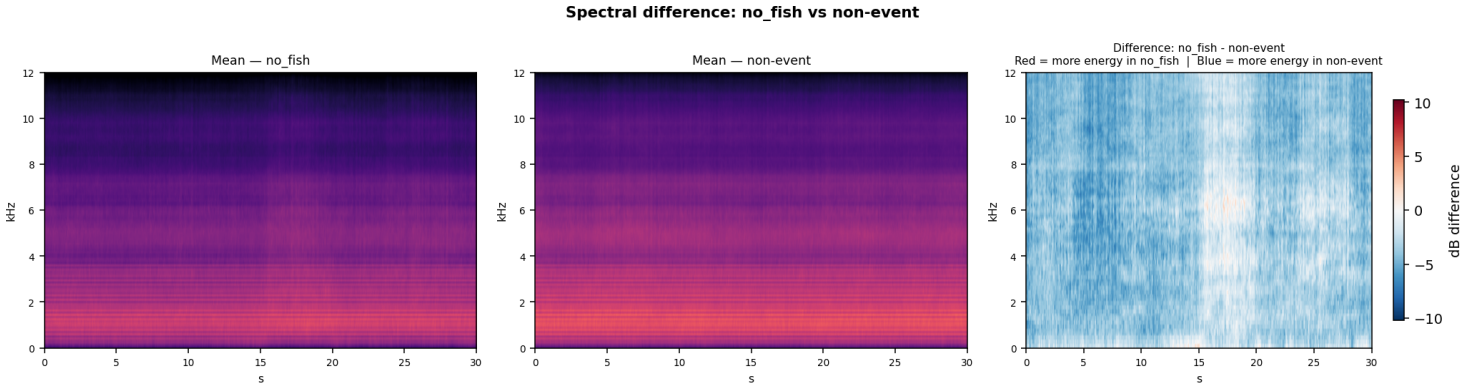

Figure 4.7: Pairwise spectral difference between no-fish and non-event classes (±10 dB scale).

4.2.2.2 Top confident 30-seconds recordings

Figure 4.8 shows the ten test recordings for which the Logistic Regression assigned the highest fish confidence. All recordings got the correct label. The spectrograms share a consistent structure: a dense broadband burst beginning around second 12–15 and persisting for several seconds. In several recordings a second burst is visible later in the clip, corresponding to chick peep response and neighbouring bird reacting to the arrival. The pre-arrival period is noticeably quieter, providing a clean contrast that makes these events easy to classify.

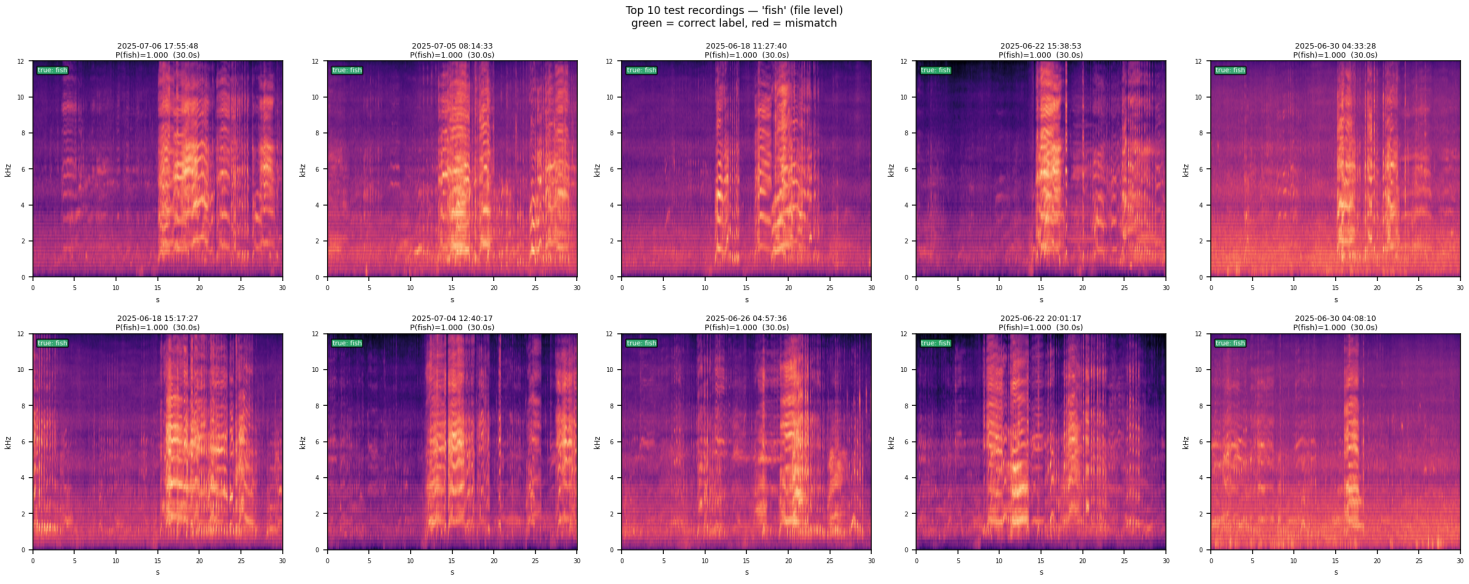

Figure 4.8: Top 10 test recordings for the fish class ranked by classifier confidence.

Figure 4.9 shows the top no-fish recordings. These are acoustically more varied than fish arrivals. Listening to the corresponding audio reveals a different vocal character: no-fish arrivals tend to elicit a quick back-and-forth exchange between the pair with little reaction from the rest of the ledge, while fish arrivals elicit a broader colony response with louder, higher-frequency calls.

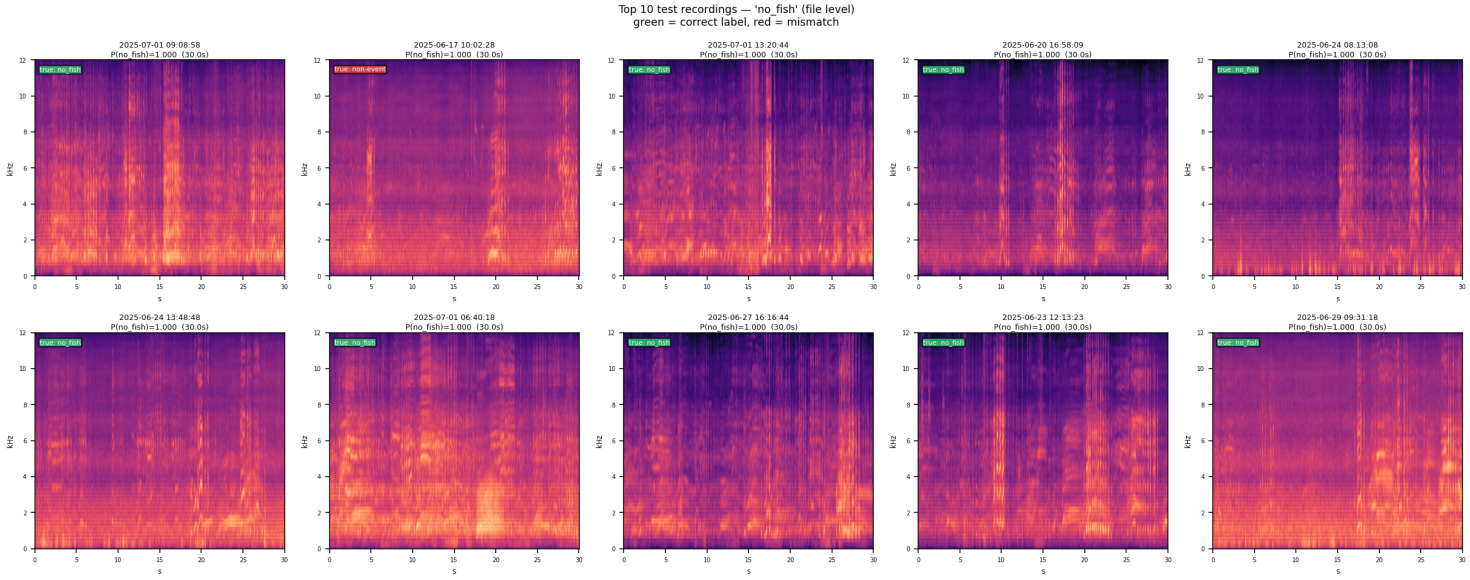

Figure 4.9: Top 10 test recordings for the no-fish class ranked by classifier confidence.

The acoustic variability of the no-fish class reflects the range of social contexts that can produce an arrival without prey. A bird returning to its breeding ledge with a partner and chick present typically elicits a brief paired exchange alongside some chick vocalisation which is acoustically similar to a fish arrival but shorter and less intense. A bird returning to its breeding ledge where a partner is present but where the egg did not hatch elicits a quieter paired exchange, without the additional vocal activity typically driven by chick presence, such as begging calls and without the strong response from the rest of the birds on the ledge.

A trespasser landing among occupied sites can elicit a vocal response from the territory owner that sounds similar to the greeting of a returning partner from a pair without a chick. If the trespasser lands at a mate-absent site, it is unlikely to elicit any response, making the event acoustically indistinguishable from a non-event. This last scenario explains the misclassification of certain no-fish arrivals as non-event, discussed further in Section 4.2.3.

Only one out of ten of the recordings in Figure 4.9 carrying a ground-truth non-event label was assigned P(no_fish) = 1 by the Logistic Regression classifier. Video inspection revealed a fight on the ledge followed by a brief paired exchange between a resident pair, a vocalisation pattern closely resembling a greeting response elicited when a trespasser arrives among occupied sites.

Figure 4.10 shows the top non-event recordings. Most are correctly labelled and show a flat, low-energy profile with no transient events. Three recordings carry a no-fish label; all involve arrivals that elicited no social response. In the first, a trespasser landed on a site that was not its own, and since all neighbouring sites were mate-absent at the time, the arrival elicited no greeting or communication between pairs. In the second and third, the bird landed on the narrow wall separating FAR3 from the adjacent ledge rather than on the breeding site itself, placing it outside the social territory of the resident pair.

In those three cases, the absence of any social trigger left the recordings acoustically empty, confirming that it is the social context of the arrival, rather than the arrival itself, that produces the acoustic signal the classifier relies on.

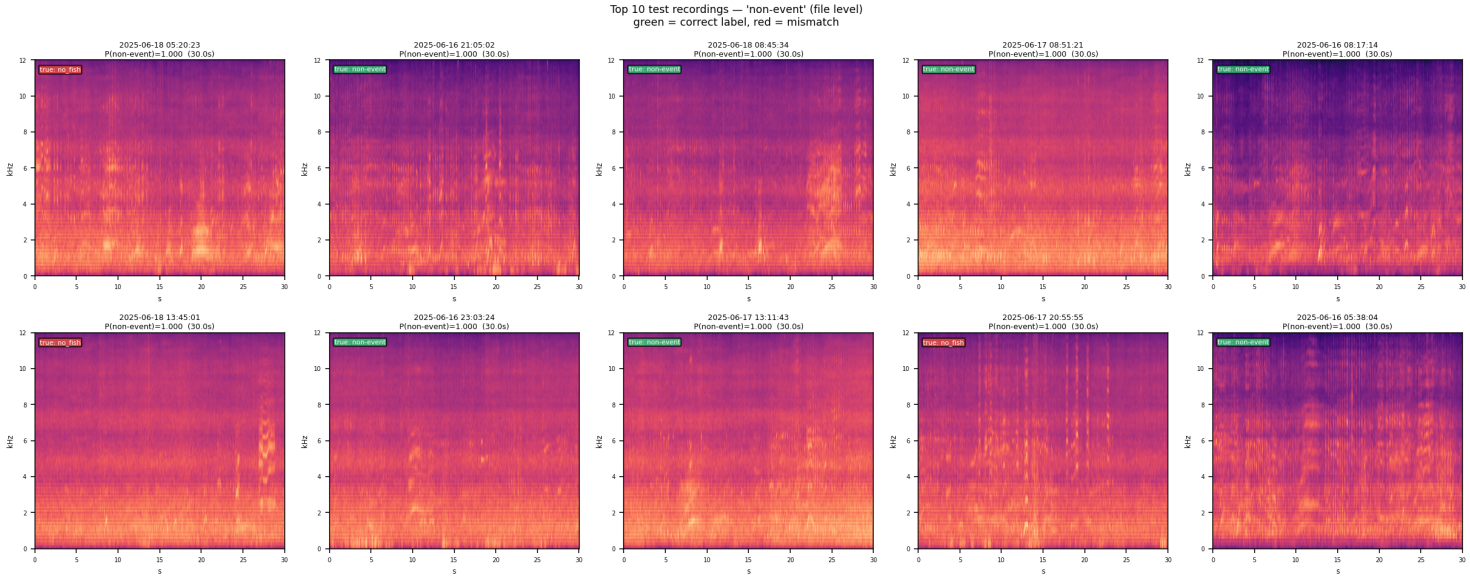

Figure 4.10: Top 10 test recordings for the non-event class ranked by classifier confidence.

4.2.2.3 Top confident 3-seconds clip

Figures 4.11, 4.12, and 4.13 show the most salient 3-second clip from each of the 10 highest-confidence 30-second recordings per class. Since BirdNET processes audio in 3-second windows but the classifier operates at file level, individual windows cannot be scored directly. Instead, the window with the highest L2 norm of the BirdNET embedding is selected from each file. This serves as a proxy for acoustic saliency in embedding space, but does not necessarily imply that the selected segment is the most discriminative for the predicted class.

Fish arrival 3-second clips (Figure 4.11) are consistently dense and broadband, with rich harmonic structure showing overlapping adult and chick vocalisations at the moment of landing. Some harmonic signal visible in several panels is characteristic of guillemot chick peep calls, with older chicks producing louder, more complex calls with visible harmonic overtones compared to the simpler peeps of younger chicks.

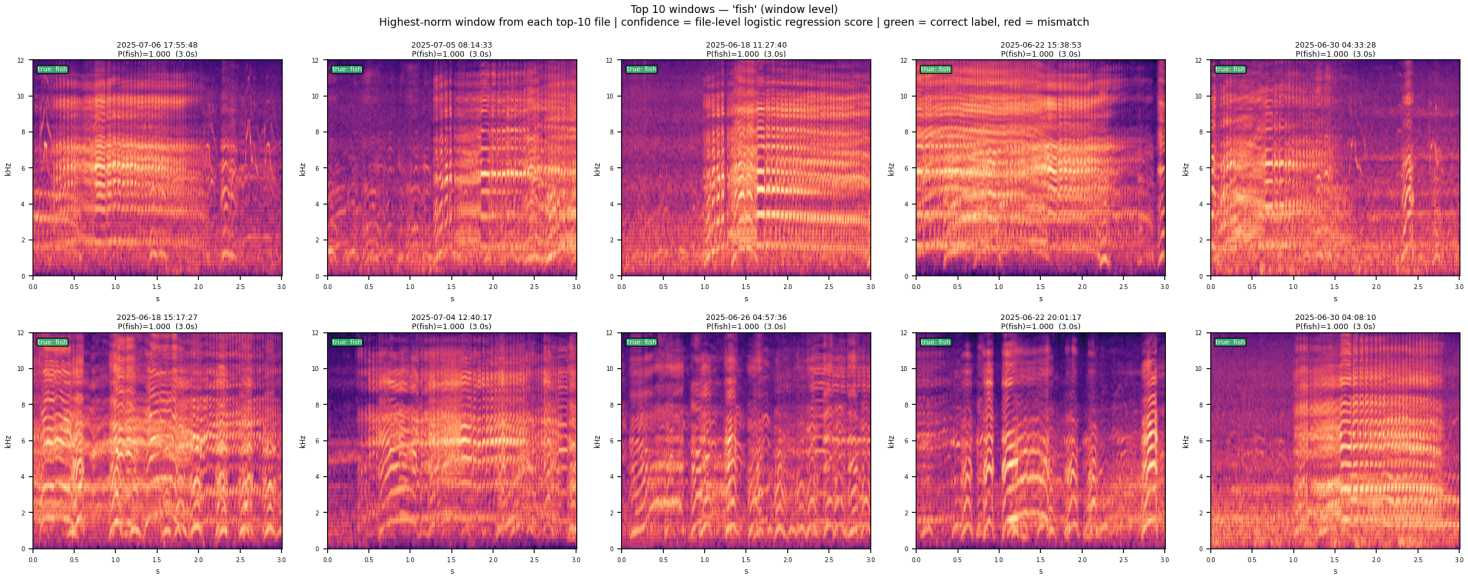

Figure 4.11: Highest-norm 3-second BirdNET window from each of the top 10 fish recordings. Windows show dense broadband energy with harmonic structure consistent with overlapping adult and chick vocalisations.

No-fish arrival 3-second windows (Figure 4.12) are also vocalisation-rich but show a different character: energy is less uniformly distributed, high-frequency content is reduced, and chick calls are largely absent. The dominant pattern is short paired exchanges between adults rather than a broad colony response. The one misclassified non-event recording shows dense low-frequency harmonic structure acoustically similar to the usual adult exchanges, explaining the confident misclassification.

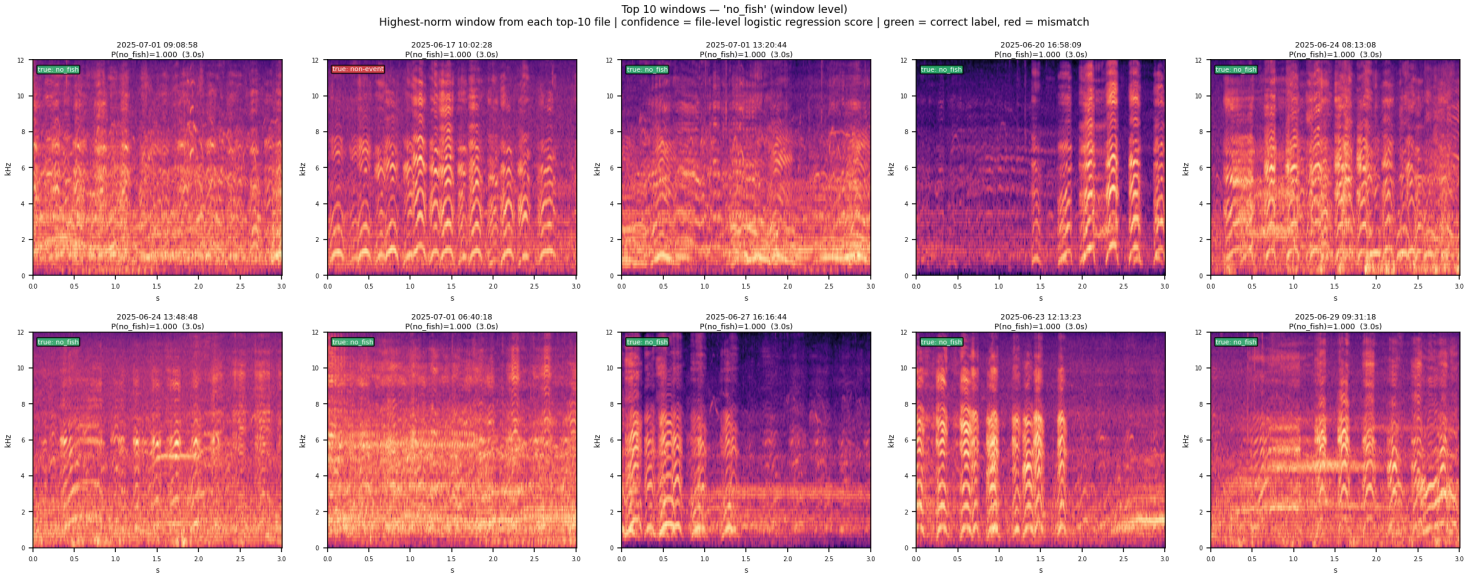

Figure 4.12: Highest-norm 3-second BirdNET window from each of the top 10 no-fish recordings. Energy is less uniformly distributed than in fish windows, with reduced high-frequency content and no visible chick calls.

Non-event windows (Figure 4.13) are the most diffuse, showing mixed broadband energy with occasional structured vocalisations but no consistent event structure. The three misclassified no-fish recordings show a flat, low-energy profile indistinguishable from background noise, confirming that the no-fish versus non-event confusion is driven by the absence of a detectable acoustic event rather than the presence of a misleading one.

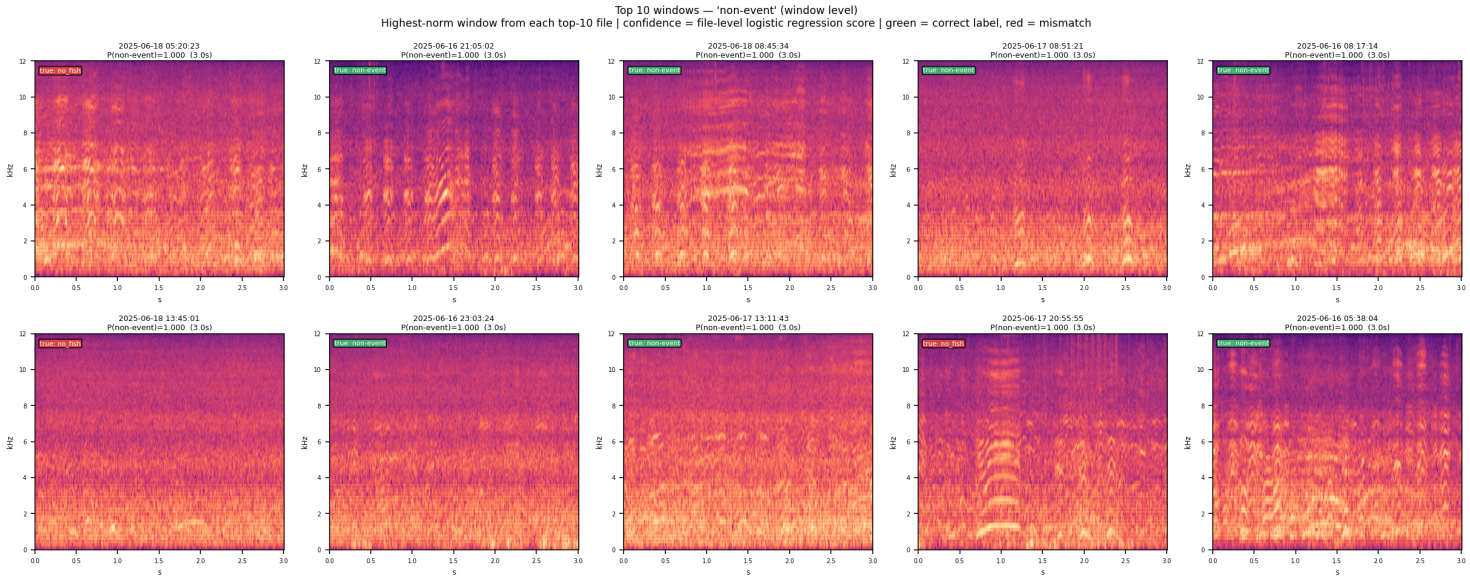

Figure 4.13: Highest-norm 3-second BirdNET window from each of the top 10 non-event recordings. Windows are diffused with no consistent event structure. The misclassified no-fish example shows a flat, low-energy profile indistinguishable from background noise.

4.2.3 Misclassification Analysis

Of the 119 test samples, 21 were misclassified by the logistic regression head trained on BirdNET embeddings (82.4% accuracy). Table 4.5 summarises the misclassification patterns grouped by error type and partner presence on the ledge.

True class	Predicted	Count	Ledge=True	Ledge=False
no_fish	non-event	8	0	8
non-event	no_fish	8	—	—
no_fish	fish	3	3	0
fish	no_fish	1	1	0
non-event	fish	1	—	—
Total		21		

Table 4.5: Misclassification summary for BirdNET with logistic regression on the manual test set. The ledge column indicates whether a partner was present on the ledge at the time of the event (True/False); non-event samples carry no ledge label (—).

All eight no-fish arrivals misclassified as non-events were birds arriving at a ledge outside their own breeding pair (`ledge=False`), while all three no-fish arrivals misclassified as fish were birds returning to their own breeding site (`ledge=True`).
The remaining errors are low-confidence or ecologically ambiguous: eight non-event samples were misclassified as no-fish, and the clip annotated *one true arrival one false and fight* was predicted as fish with 0.97 confidence despite its no-fish label.

4.2.4 Effect of Weather Features on Classification

Weather features were added to the BirdNET embeddings to test whether environmental data improves classification performance with the MLP classifier. After alignment and median imputation, a stability-based selection procedure was applied, retaining the most consistently informative features across bootstrap runs. These were mainly temperature- and humidity-related variables as can be seen in Figure 4.14 which shows the stability-ranked weather features.
Across different values of k (1–9 selected features), performance remained stable, with only small variations. The best result was obtained using 4 features, giving the highest overall performance.
Table 4.6 compares the best weather-augmented model with the BirdNET-only baseline. Overall, the inclusion of weather features yields improvement, mainly driven by improved separation of the non-event class.

Model	Accuracy	Precision	Recall
With weather (best $k = 4$)	0.8319	0.8433	0.8403
BirdNET only (baseline)	0.8067	0.8163	0.8170

Table 4.6: Overall performance comparison between BirdNET-only and weather-augmented models.

Per-class results are shown in Tables 4.7 and 4.8. The no-fish class remains the weakest across both settings, while fish is consistently well detected and non-event benefits most from the added environmental context.

Class	Precision	Recall
fish	0.886	0.939
no_fish	0.875	0.651
non-event	0.769	0.930

Table 4.7: Per-class performance for the best weather-augmented model ($k = 4$).

Class	Precision	Recall
fish	0.886	0.939
no_fish	0.818	0.628
non-event	0.745	0.884

Table 4.8: Per-class performance for the BirdNET-only baseline model.

4.2.4.1 Selected weather features

The stability-ranked weather features are shown in Figure 4.14.

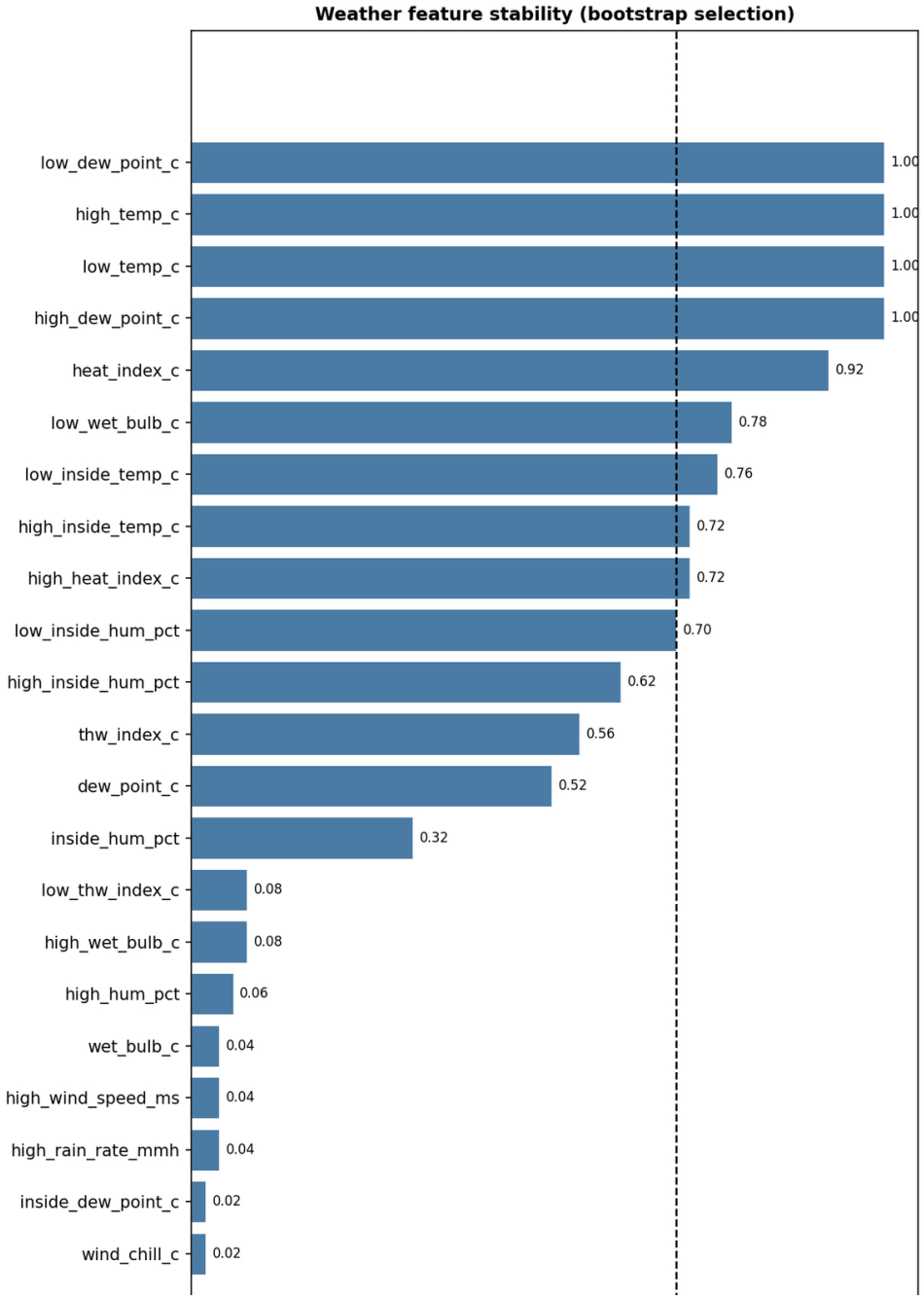

Figure 4.14: Stability-based ranking of weather features across bootstrap runs. The vertical dotted line shows a confidence threshold (70% selection frequency).

The four most stable features: `low_dew_point_c`, `high_temp_c`, `low_temp_c`, and `high_dew_point_c` all achieved a perfect selection frequency of 1.0 across bootstrap runs, meaning they were chosen in every resample. All four relate to temperature and humidity conditions, suggesting these are the most consistently informative environmental variables for classification.

This pattern is consistent with known behavioural responses of guillemots to heat stress, which can alter colony activity levels and vocal behaviour under warmer or more humid conditions. Temperature and dew point may therefore act as proxies for thermally-driven changes in colony acoustic activity. The gain from weather augmentation thus appears to be driven by a small set of highly stable environmental variables that capture this thermally-linked variation.

4.2.5 Generalisation Beyond the Ground-truth Dataset

To assess how well the classifier generalises beyond the manually curated dataset, two evaluation setups were compared across three datasets: the manual FAR3 dataset, and two automatically labelled datasets derived from different versions of the YOLO detection model. The original automatic dataset was created from an earlier YOLO model run, and served as the basis from which the manual dataset was constructed. The updated automatic dataset was generated from a newer YOLO model run on the same ledge, allowing us to assess whether improved detections translate to better automatic labels. In the first evaluation setup, a logistic regression head was fitted on embeddings from each automatic dataset and tested both on that same dataset and on the manual FAR3 test set. In the second, the logistic regression head fitted on the manual FAR3 dataset, was tested on the original automatic dataset and then manual FAR3 set. This comparison isolates the effect of training label quality on classification performance. Results are presented in Table 4.9.

		Tested on auto			Tested on manual		
Fitted on	Class	Precision	Recall	F1	Precision	Recall	F1
			original				
Manual FAR3	fish	0.73	0.64	0.68	0.89	0.97	0.93
	no_fish	0.44	0.72	0.55	0.75	0.70	0.72
	non-event	0.43	0.21	0.29	0.77	0.77	0.77
			updated				
Auto FAR3 (updated)	fish	0.84	0.71	0.77	0.70	1.00	0.82
	no_fish	0.51	0.51	0.51	0.45	0.47	0.46
	non-event	0.52	0.60	0.56	0.57	0.37	0.45
			original				
Auto FAR3 (original)	fish	0.68	0.57	0.62	0.72	0.94	0.82
	no_fish	0.35	0.36	0.35	0.67	0.51	0.58
	non-event	0.42	0.47	0.44	0.70	0.70	0.70

Table 4.9: BirdNET logistic regression per-class performance, comparing classifiers fitted on manual versus automatic labels. The set line names the automatic test set used; the manual block always uses the manual FAR3 test set.

Across all conditions, fish arrivals achieved the highest F1-score. For the model fitted on manual labels, testing on the manual set rather than the original automatic set raised fish F1 from 0.68 to 0.93, no-fish from 0.55 to 0.72, and non-event from 0.29 to 0.77, indicating that incorrect labels in the original automatic dataset are the

primary driver of the performance gap. The same effect appears for the model fitted on the original automatic labels, where moving from the automatic to the manual test set raised fish F1 from 0.62 to 0.82, no-fish from 0.35 to 0.58, and non-event from 0.44 to 0.70. This is also expected, since the manual dataset was curated from the original automatic dataset, so the two share most of their events and differ mainly in label quality. The model fitted on the updated automatic labels did not follow this pattern: testing on the manual set rather than the updated automatic set raised fish F1 (0.77 to 0.82) but lowered no-fish (0.51 to 0.46) and non-event (0.56 to 0.45). This reflects a tendency to over-predict fish: on the manual test set the model reached a fish recall of 1.00 at a precision of 0.70, so it captured every fish arrival but drew some no-fish and non-event samples into the fish class, lowering their recall.

Of the 441 automatically detected fish events in the updated YOLO dataset, 250 (56.7%) corresponded to a manually curated event, and of the 465 in the original YOLO dataset, 277 (59.6%) corresponded to a manually curated event. Of the 46 manual fish events absent from the original automatic dataset, 11 (24%) had no fish detection in the YOLO database during the corresponding time window.

The large proportion of automatically detected events without a corresponding manual annotation suggests that many automatic labels are false positives, which are penalised as classifier errors regardless of the acoustic signal. The no-fish class is likely the most affected, as false positive fish detections from YOLO directly corrupt the no-fish labels of neighbouring events. Taken together, these results suggest that the quality of the automatic labelling pipeline is a key bottleneck, and that the acoustic classifier may generalise better than the raw numbers indicate.

4.2.6 Cross-Site Generalisation to Another Ledge: BONDEN6

To evaluate cross-site generalisation, two conditions were compared on the BONDEN6 test set (n = 98: fish = 33, no-fish = 33, non-event = 32), recorded at a different ledge from the FAR3 training site. In the first, a logistic regression head was fitted on BONDEN6 embeddings. In the second, a logistic regression head fitted on the manual FAR3 dataset was tested directly on BONDEN6. Per-class results are reported in Table 4.10 and confusion matrices are shown in Figure 4.15.

Tested on	Fitted on BONDEN6			Fitted on manual FAR3		
BONDEN6	Precision	Recall	F1	Precision	Recall	F1
fish	0.68	0.70	0.69	0.76	0.67	0.71
no_fish	0.54	0.39	0.46	0.48	0.67	0.56
non-event	0.57	0.72	0.64	0.52	0.38	0.44
Accuracy		0.60			0.57	

Table 4.10: BirdNET logistic regression per-class performance on the BONDEN6 test set, comparing a classifier fitted on BONDEN6 versus one fitted on the manual FAR3 dataset.

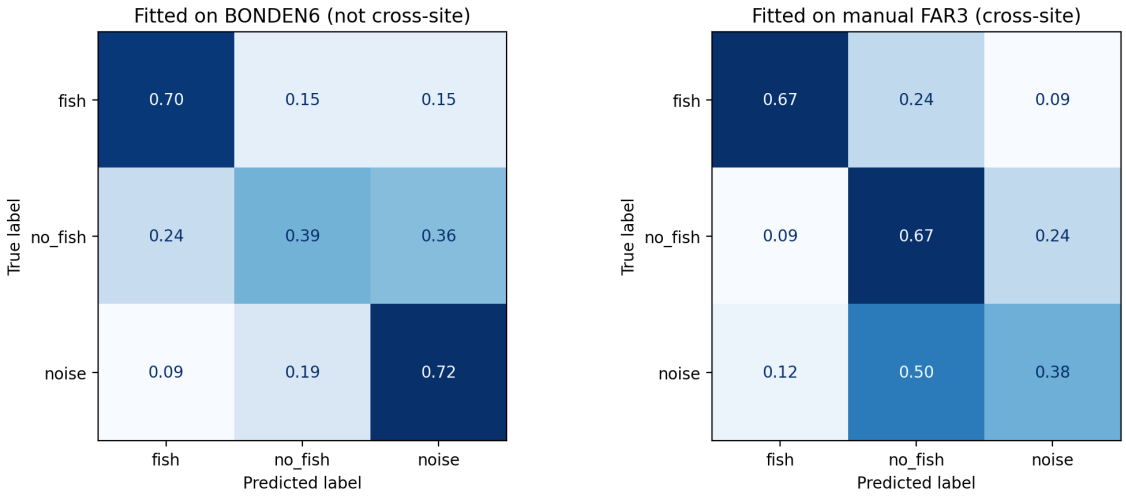

Figure 4.15: Normalised confusion matrices on the BONDEN6 test set, for a classifier fitted on BONDEN6 (left) and one fitted on the manual FAR3 dataset (right).

The manual-labels condition improves fish and no-fish detection relative to the auto-labels condition, with fish F1 rising from 0.69 to 0.71 and no-fish from 0.46 to 0.56. Non-event performance decreases (0.64 to 0.44), resulting in a slightly lower overall accuracy (0.60 vs 0.57). The improvement on no-fish is the most notable finding, a classifier fitted on clean manual annotations from a different ledge detects no-fish arrivals at BONDEN6 more reliably than one fitted on BONDEN6's own automatic labels, suggesting that the manual dataset captures a more generalisable acoustic boundary for this class.

The fish class transfers most reliably across sites under both conditions, consistent with fish arrivals producing a distinctive and acoustically consistent broadband response regardless of ledge. The no-fish and non-event classes remain harder to distinguish, which is expected given that the acoustic characteristics of these classes depends on local social context and background conditions that may vary between ledges. The relatively small size of the BONDEN6 test set means these results should be interpreted with caution, but the overall pattern is encouraging for the prospect of deploying a classifier trained on manual FAR3 data across multiple ledges without site-specific retraining.

5

Conclusions

This thesis presented a machine listening pipeline for classifying guillemot arrival events at the AukLab on Stora Karlsö, distinguishing three acoustic classes: arrivals with fish, arrivals without fish, and non-events (background colony noise). The pipeline combined event-driven labelling from YOLO-based video detections with frozen pre-trained CNN embeddings and lightweight classifier heads, trained and evaluated on a manually curated dataset of 1159 samples from the FAR3 ledge during the 2025 breeding season, demonstrating that arrival events can be automatically detected and classified from acoustic recordings using video-derived annotations (RQ1).

Of the three backbone models evaluated, BirdNET, PANN, and Perch, BirdNET achieved the strongest performance, with an overall accuracy of 80.7% and a macro F1-score of 0.82 on the held-out test set. PANN performed slightly worse at 78.2%, while Perch lagged considerably at 64.7%. UMAP projections confirmed that BirdNET embeddings produced the clearest class separation of the three models, with fish arrivals forming a relatively compact region and a visible gradient from non-event to no-fish to fish suggesting a meaningful ordering of acoustic similarity. No-fish, however, overlapped with both other classes, reflecting its intermediate acoustic character. The choice of classifier head had limited impact on overall performance, with logistic regression, SVM, and MLP all achieving comparable accuracy, suggesting that the discriminative information is largely encoded in the BirdNET embeddings themselves which shows that CNN-based bioacoustic models can perform reliably under noisy, real-world colony conditions (RQ5).

Across all models, fish arrivals were consistently the most acoustically separable class, driven by a broadband energy burst centred at the moment of landing, a response reflecting simultaneous adult and chick vocalisations. The no-fish class proved the most challenging: difference maps and misclassification analysis confirmed that arrivals without fish are acoustically subdued, with fewer chick peep calls and weaker adult vocalisations, a pattern that directly characterises how guillemot vocalisations differ between the two arrival types (RQ2). This ambiguity is ecologically genuine rather than a modelling artefact, as a subset of no-fish arrivals generated no discernible acoustic event and are effectively indistinguishable from non-events at the recording level.

Misclassification analysis revealed an ecologically interpretable pattern: all eight no-fish arrivals misclassified as non-events corresponded to birds returning to a ledge other than their own breeding site, where chick-related vocal interactions would be absent. This finding suggests that the model implicitly learns the social context of arrivals, and that ledge-specific acoustic signatures play a meaningful role in the

classification. Incorporating weather features alongside BirdNET embeddings produced a marginal improvement in accuracy 83.2%, with the most stable predictors corresponding to humidity and temperature, though the modest gain suggests that acoustic features carry most of the discriminative signal under the conditions studied (RQ3).

The pipeline generalised beyond the manually curated training data, achieving an F1-score of 0.82 for fish arrivals when trained on the automatically constructed FAR3 dataset and tested on the manual set, and 0.71 on the held-out BONDEN6 ledge. Performance on the no-fish and non-event classes was more sensitive across datasets, generally improving when evaluated against the clean manual labels rather than the automatic ones. This is consistent with the higher acoustic ambiguity of those classes and differences in label quality between the manual and automatic annotation pipelines. The YOLO model version used for automatic event extraction had a measurable impact on classification performance, underscoring the importance of detection quality in event-driven annotation workflows.

Taken together, these results demonstrate that pre-trained bioacoustic embeddings, combined with event-driven labelling from synchronised video, enable reliable and scalable acoustic classification of prey delivery events in a dense seabird colony. The approach reduces reliance on manual annotation while retaining ecologically interpretable structure in the learned representations (RQ4), providing a practical foundation for long-term, automated monitoring of guillemot provisioning behaviour at Stora Karlsö and contributing to the broader effort of developing non-invasive tools for large-scale ecological observation of seabird populations.

6
Future Work

Several directions could extend this work. On the acoustic side, noise removal could be improved by exploiting the eight-channel DPA microphone array to spatially isolate sounds originating from the target ledge, and additional audio processing could further improve classifier robustness. Testing different audio window lengths would also help determine whether shorter, event-centred clips yield better classification accuracy.

Beyond BirdNET, other audio foundation models such as PERCH and PANN warrant further exploration. While these were tested in this work, a more thorough investigation using alternative classifier heads beyond the MLP and logistic regression, combined with more extensive hyperparameter searches, could yield meaningful improvements. Similarly, the effect of different train/validation/test splits should be examined more carefully, as the relatively small dataset size means performance estimates may be sensitive to how data is partitioned.

On the visual side, improving the fish detection model either by fine-tuning YOLO on more colony data or by exploring alternatives such as SAM (Segment Anything Model) could enable automatic dataset construction, removing the need for manual curation and making it easier to scale to new ledges and seasons. Extending the pipeline to additional ledges beyond FAR3 and BONDEN6 would further test how well the model generalises across the colony.

Finally, from a biological perspective, it would be interesting to investigate acoustic differences between pairs that still have eggs and those that have lost them. Such differences could potentially provide a non-invasive indicator of reproductive status and changes in colony behaviour.

Bibliography

[1] J. Hentati-Sundberg et al., "Technological Evolution Generates New Answers and New Ways Forward: A Progress Report from the First Decade at the Karlsö Auk Lab," *Marine Ornithology*, vol. 53, no. 1, Apr. 2025, ISSN: 1018-3337, 2074-1235. DOI: 10.5038/2074-1235.53.1.1612. Accessed: May 8, 2026. [Online]. Available: https://digitalcommons.usf.edu/marine_ornithology/vol53/iss1/7.

[2] J. H. Andersen et al., "Long-term temporal and spatial trends in eutrophication status of the Baltic Sea: Eutrophication in the Baltic Sea," en, *Biological Reviews*, vol. 92, no. 1, pp. 135–149, Feb. 2017, ISSN: 14647931. DOI: 10.1111/brv.12221. Accessed: Mar. 5, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/brv.12221.

[3] A. Kershenbaum et al., "Automatic detection for bioacoustic research: A practical guide from and for biologists and computer scientists," en, *Biological Reviews*, vol. 100, no. 2, pp. 620–646, Apr. 2025, ISSN: 1464-7931, 1469-185X. DOI: 10.1111/brv.13155. Accessed: Feb. 3, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/brv.13155.

[4] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," 2021, Version Number: 1. DOI: 10.48550/ARXIV.2112.06725. Accessed: Jan. 19, 2026. [Online]. Available: https://arxiv.org/abs/2112.06725.

[5] J. Podos and M. S. Webster, "Ecology and evolution of bird sounds," en, *Current Biology*, vol. 32, no. 20, R1100–R1104, Oct. 2022, ISSN: 09609822. DOI: 10.1016/j.cub.2022.07.073. Accessed: Jan. 28, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960982222012258.

[6] A. Márquez-Rodríguez et al., "A bird song detector for improving bird identification through deep learning: A case study from Doñana," en, *Ecological Informatics*, vol. 90, p. 103 254, Dec. 2025, ISSN: 15749541. DOI: 10.1016/j.ecoinf.2025.103254. Accessed: Jan. 30, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954125002638.

[7] R. Schwinger et al., *Foundation Models for Bioacoustics – a Comparative Review*, Version Number: 1, 2025. DOI: 10.48550/ARXIV.2508.01277. Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/abs/2508.01277.

[8] D. Stowell, *Can deep learning understand animal sounds? State of the art in 2023*, May 2023. DOI: https://bioacousticai.eu/can-deep-learning-understand-animal-sounds-state-of-the-art-in-2023/.

[9] K. McGinn, S. Kahl, M. Z. Peery, H. Klinck, and C. M. Wood, "Feature embeddings from the BirdNET algorithm provide insights into avian ecology,"

en, *Ecological Informatics*, vol. 74, p. 101 995, May 2023, ISSN: 15749541. DOI: 10.1016/j.ecoinf.2023.101995. Accessed: Feb. 6, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954123000249.

[10] S. Allen-Ankins, S. Hoefer, J. Bartholomew, S. Brodie, and L. Schwarzkopf, "The use of BirdNET embeddings as a fast solution to find novel sound classes in audio recordings," *Frontiers in Ecology and Evolution*, vol. 12, p. 1 409 407, Jan. 2025, ISSN: 2296-701X. DOI: 10.3389/fevo.2024.1409407. Accessed: Feb. 6, 2026. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fevo.2024.1409407/full.

[11] B. Ghani, V. J. Kalkman, B. Planqué, W.-P. Vellinga, L. Gill, and D. Stowell, "Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification," en, *Scientific Reports*, vol. 15, no. 1, p. 16 273, May 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-00996-2. Accessed: Jan. 28, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-00996-2.

[12] L. Rauch, R. Heinrich, I. Moummad, A. Joly, B. Sick, and C. Scholz, *Can Masked Autoencoders Also Listen to Birds?* Version Number: 4, 2025. DOI: 10.48550/ARXIV.2504.12880. Accessed: Jan. 27, 2026. [Online]. Available: https://arxiv.org/abs/2504.12880.

[13] A. Gascoyne and W. Lomas, "First-of-its-kind AI model for bioacoustic detection using a lightweight associative memory Hopfield neural network," en, *Ecological Informatics*, vol. 91, p. 103 382, Nov. 2025, ISSN: 15749541. DOI: 10.1016/j.ecoinf.2025.103382. Accessed: Jan. 22, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954125003917.

[14] C. H. S. Prasanna, *A Comprehensive Survey on Deep Learning Techniques for Bird Species Classification Using Image and Audio Modalities*, Oct. 2025. DOI: 10.36227/techrxiv.176148812.28724055/v1. Accessed: Jan. 28, 2026. [Online]. Available: https://www.techrxiv.org/users/981164/articles/1347255-a-comprehensive-survey-on-deep-learning-techniques-for-bird-species-classification-using-image-and-audio-modalities?commit=a98f1b89ad8281eed835df971057f87518fb11df.

[15] V. Tennekoon et al., "Multi-Modal Behavior Classification Using Audio Visual Data of Vanellus Indicus," in *2024 6th International Conference on Advancements in Computing (ICAC)*, Colombo, Sri Lanka: IEEE, Dec. 2024, pp. 133–138, ISBN: 979-8-3315-1787-8. DOI: 10.1109/ICAC64487.2024.10851153. Accessed: Feb. 5, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/10851153/.

[16] Z. Miao et al., "Multi-modal Language models in bioacoustics with zero-shot transfer: A case study," en, *Scientific Reports*, vol. 15, no. 1, p. 7242, Feb. 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-89153-3. Accessed: Jan. 27, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-89153-3.

[17] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," en, *Ecological Informatics*, vol. 61, p. 101 236, Mar. 2021, ISSN: 15749541. DOI: 10.1016/j.ecoinf.2021.101236.

Accessed: Jan. 27, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954121000273.

[18] B. Van Merriënboer, J. Hamer, V. Dumoulin, E. Triantafillou, and T. Denton, "Birds, bats and beyond: Evaluating generalization in bioacoustics models," *Frontiers in Bird Science*, vol. 3, p. 1 369 756, Jul. 2024, ISSN: 2813-3870. DOI: 10.3389/fbirs.2024.1369756. Accessed: Jan. 27, 2026. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbirs.2024.1369756/full.

[19] R. R. K. V. S. N., J. Montgomery, S. Garg, and M. Charleston, "Bioacoustics Data Analysis – A Taxonomy, Survey and Open Challenges," *IEEE Access*, vol. 8, pp. 57 684–57 708, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2978547. Accessed: May 10, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/9025054/.

[20] D. G. Ainley, D. N. Nettleship, and A. E. Storey, "Common Murre (Uria aalge)," en, in *Birds of the World*, S. M. Billerman, B. K. Keeney, P. G. Rodewald, and T. S. Schulenberg, Eds., Cornell Lab of Ornithology, Aug. 2021. DOI: 10.2173/bow.commur.02. Accessed: May 18, 2026. [Online]. Available: https://birdsoftheworld.org/bow/species/commur/2.0/introduction.

[21] T. Lengagne, M. P. Harris, S. Wanless, and P. J. Slater, "Finding your mate in a seabird colony: Contrasting strategies of the Guillemot *Uria aalge* and King Penguin *Aptenodytes patagonicus*," en, *Bird Study*, vol. 51, no. 1, pp. 25–33, Mar. 2004, ISSN: 0006-3657, 1944-6705. DOI: 10.1080/00063650409461329. Accessed: Apr. 8, 2026. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00063650409461329.

[22] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, *YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection*, Version Number: 5, 2025. DOI: 10.48550/ARXIV.2509.25164. Accessed: May 20, 2026. [Online]. Available: https://arxiv.org/abs/2509.25164.

[23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, *PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition*, Version Number: 5, 2019. DOI: 10.48550/ARXIV.1912.10211. Accessed: Mar. 25, 2026. [Online]. Available: https://arxiv.org/abs/1912.10211.

[24] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, *Perch 2.0: The Bittern Lesson for Bioacoustics*, Version Number: 2, 2025. DOI: 10.48550/ARXIV.2508.04665. Accessed: Jan. 28, 2026. [Online]. Available: https://arxiv.org/abs/2508.04665.

DEPARTMENT OF PHYSICS - UNIVERSITY OF GOTHENBURG
DEPARTMENT OF PHYSICS AND ASTRONOMY - CHALMERS
Gothenburg, Sweden
www.chalmers.se www.gu.se

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY